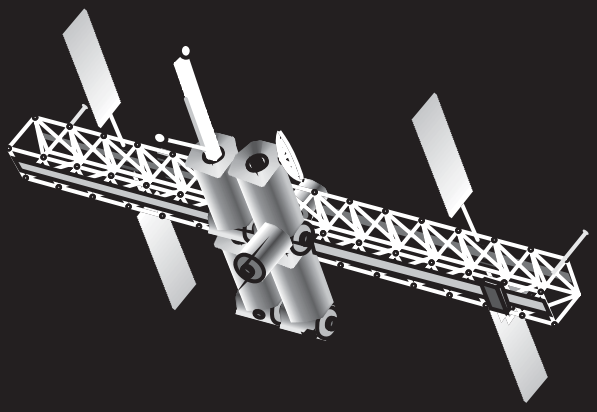
ENSC 803
Simon Fraser University
School of Engineering Science

2008 Course Proceedings



Graduate Communication:

Applying Strategies of Writing for Publication in Technical Journals

A 2-stage Reference Picture Selection H.264 Encoder and a Novel Non-Casual Whole Frame Concealment Algorithm

S. Mohsen Amiri, Ivan V. Bajić

***Abstract*—Since compressed bit-streams are subject to corruption, or are lost, error control techniques are very important for video communication, and video communication systems could confront errors in two ways, which are error resilience, and error concealment. Because the compressed video bit-streams are very sensitive to packet loss, the error easily propagates through other frames, and the video communication system should have an error resilience capability to avert this problem. Reference Picture Selection (RPS) is a powerful error resilience technique for video streaming, yet it consumes significant computation resources at the time of streaming. The first part of this paper proposes a two-stage H.264 compatible encoder that enables the use of RPS at significantly reduced computational cost and with only a minor penalty in coding performance. The second part of the paper proposes a non-causal whole-frame concealment algorithm. Future, the second part of the paper proposes a non-causal whole-frame concealment algorithm. This algorithm tries to conceal lost frames using information from three different source: information from previous frames, upcoming frames, and the reset frame (if the communication system uses NACK-RPS as the error resilience method). To the best knowledge of the author, all previously proposed methods have tried to use the information offered by previously frames to conceal the whole-frame loss.**

***Index Terms*—Video coding, Video streaming, Reference Picture Selection, Error concealment, Error resilience, and Whole-frame loss.**

I. Introduction

RECENTLY, video communication over lossy channels, such as the internet and wireless channels, has received increasing attention. During transmission over a lossy channel, the communication system could confront errors in two ways: error resilience and error concealment.

First, error propagation is a serious problem affecting transmission of predicatively coded video. Errors cause the received frame at the decoder to be different from the ones used in the encoder for motion compensation. Once an erroneous frame enters the decoder's motion compensation loop, the error propagates to the future frames and degrades the quality of the decoded video.

To prevent such error propagation, several methods have been proposed [1], most relying on the feedback information about errors sent from the decoder to the encoder. In the simplest such error control scheme, upon receiving a notification about an error, the encoder encodes the following frame in the I-mode. Since the I-frame resets the state of the coding loop, error propagation is stopped. However, compression efficiency suffers significantly, since intra-frame coding is much less efficient than motion-compensated predictive coding. An alternative method for stopping error propagation, but without significant loss of coding efficiency, is Reference Picture Selection (RPS) [2]. In this scheme, the encoder will not use for prediction those frames that have been flagged as erroneous by the decoder. In this way, error propagation is stopped without significant loss of coding efficiency, because the encoder can still remove some temporal redundancy form the video signal through motion compensation. For this reason, provisions for enabling RPS have been incorporated in Annex U of H.263 video coding standard, and also in the latest H.246/AVC video coding standard [3].

Simulations of RPS have demonstrated its effectiveness and advantages against other popular error control schemes such as Forward Error Correction (FEC) and Automatic Repeat reQuest (ARQ) [4]. However, the use of RPS in practical video streaming systems has been fairly limited. The main reason is the computational burden associated with RPS at streaming time. The frame to be used as the reference for motion-compensated prediction is not known in advance, and needs to be chosen at transmission time based on decoder feedback. Hence, the encoder needs to operate in real time. In this paper we propose a H.264-compatible two-stage encoder that significantly reduces the computational cost of RPS at the expense of additional disk space at the server. Its structure is described in the following section.

The second improvement involves trying to estimate the lost frame in the decoder and adding an estimated one in order to enhance the decoded video (error concealment). [5], [6], [7], and [8] proposed several error concealment methods,which use temporal, spatial, and spatiotemporal interpolation methods. [9], and [10] provide a good review of different error concealment methods. [6] uses spatial information of the lost MB in the Discrete Cosine Transform (DCT) domain and assumes the recovered area is smooth. [7], and [8] uses spatiotemporal information and uses a boundary matching algorithm[11], [12] to reestimate a motion vector for the lost MB.

But when a whole-frame loss happens, none of above described methods are applicable because no neighborhood is available to use in interpolation or to use in boundary matching algorithm. [13] proposed a whole-frame loss concealment method, which uses an optical flow (OF) model [14]

S. Mohsen Amiri is with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A1S6 e-mail: samiri@sfu.ca

Ivan V. Bajić is with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A1S6 e-mail: ibajic@sfu.ca

Manuscript received March , 2008; revised April 3, 2008.

to estimate the motion vectors for the lost frame and uses zero residual to complete the frame concealment. Here, the proposed method in [13] assumes that every object has a constant velocity and performs the following steps [15]:

1) **Estimation of the optical flow:** If frame *t* has been lost, for each pixel in frame *t-1*, the algorithm computes a forward motion vector (FMV) in order to fill up a FMV field by assuming each pixel had a constant velocity. Moreover, to save memory, it rounds these motion vectors into half pixel resolution.
2) **Spatial regularization of the FMV:** Since the encoder can code some MBs in intra mode, some empty areas are available in the FMV field. The algorithm fills them up via vector median filter with a 7×7 kernel size and then uses another vector median filter with 15×15 kernel for regularization of the FMV field.
3) **Projection onto missing frame:** The algorithm reconstructs frame t using the FMV field generated in the previous step. Since the motion-vectors have half pixel accuracy, the algorithm should copy every pixel in frame *t-1* to a 2×2 neighborhood in frame *t*, onto the position it points to according to associated FMV. Because some pixels of the reconstructed image can have multiple contributions, the algorithm simply average them.
4) **Interpolation of missing pixels:** After projection, the reconstructed frame may have some empty areas. The algorithm fills these points by recursively using a 9×9 median filter.
5) **Filtering and downsampling:** As a final step, the reconstructed frame is created by downsampling of the image from the previous step with factor 2

Despite the high performance of the proposed algorithm in [13], it consumes significant computational resources in the decoder because it works with pixels. In [15], the authors proposed a block-based error concealment method for H.264/AVC, which works as outlined:

1) **Searching for a suitable reference frame:** Since H.264/AVC encoder can support multiple reference frames, the proposed algorithm in this paper searches for the nearest available picture that has significant number of inter coded MBs, which is usually frame *t-1* if frame *t* is the lost one.
2) **Motion vector projection:** For each pixel in frame *t-1*, the algorithm projects its motion vector to frame *t* and adds this information to its corresponding 4×4 block in frame *t*.
3) **Statistics collection:** As well as the information stored in 4×4 blocks, it computes the following statistics for each 16×16 in frame *t*.
   - $M_{16\times16}$: Average of projected MVs into a 16×16 MB
   - $M_{4\times4}$: Average of projected MVs into a 4×4 block
   - $\sigma_{16\times16}$: Variance of projected MVs into a 16×16 MB
   - $\sigma_{4\times4}$: Variance of projected MVs into a 4×4 block
   - $N_{16\times16}$: Number of projected MVs into a 16×16 MB

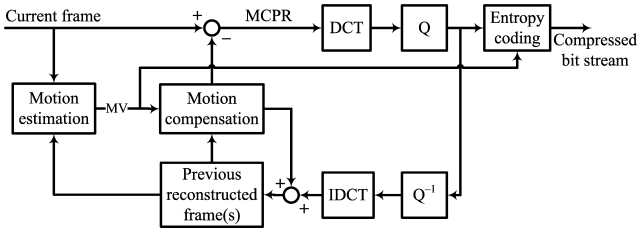


Fig. 1. Standard video coder based on motion compensation.

   - $N_{4\times4}$: Number of projected MVs into a 4×4 block
4) **MB and block level motion estimation:** In this level, if the following condition is true,

$$(M_{16\times16} < THR_{V_{16\times16}})AND(N_{16\times16} > THR_{N_{16\times16}})$$

, the algorithm decides to assign $M_{16\times16}$ to the corresponding MB as the motion vector. If the algorithm fails in the previous step it checks the following condition:

$$(\sigma_{4\times4} < THR_{V_{4\times4}})AND(N_{4\times4} > THR_{N_{4\times4}})$$

. If it is true, the algorithm assigns $M_{4\times4}$ to the corresponding block. This step is iterated, lowering the threshold value, until all the blocks have their corresponding motion vector.
5) **Picture reconstruction:** Finally, enough information is gathered to reconstruct the lost frame in the motion compensation loop and the loop filter available in the encoder.

The rest of this paper is organized as follows. We describe the proposed 2-stage encoder in section II and non-causal whole-frame concealment algorithm in section III. Then we present the experimental results in section IV to verify the performance of proposed encoder and the non-causal algorithm. Finally in section V, we summarize proposed 2-satege encoder and concealment algorithm.

II. Proposed 2-stage encoder

Fig. 1 shows the structure of a standard video coder based on motion-compensated prediction [14]. Incoming video frame (called "current frame") is subject to motion estimation with respect to previous reconstructed frame(s). Using the resulting motion vectors (MVs), a prediction of the current frame is made. The difference between the current frame and the prediction (called motion-compensated prediction residual, MCPR) is then subject to Discrete Cosine Transform (DCT), quantization, and entropy coding along with MVs, to create the compressed bit stream. In addition, quantized MCPR is dequantized, transformed with inverse DCT (IDCT) and added to the prediction to create the reconstructed version of the current frame, which is then stored in the frame buffer for future motion estimation and compensation.

According to [16], [17], motion estimation accounts for about 80% of the processing time in P-frame coding. However, in typical streaming applications such as video-on-demand, video content is available ahead of streaming time. This gives us the opportunity to perform motion estimation ahead of

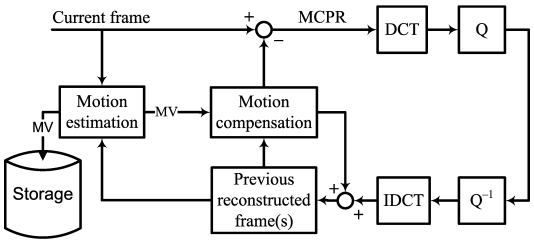


Fig. 2. The first stage of the two-stage encoder.

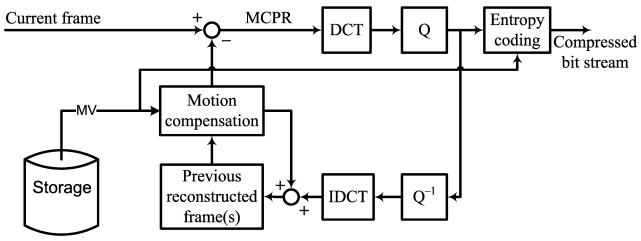


Fig. 3. The second stage of the two-stage encoder.

time, leaving motion compensation, quantization, and entropy coding for real-time operation during the streaming session. Hence, in our proposed two-stage encoder, we reduce real-time processing requirements of the video encoder by decoupling motion estimation from motion compensation.

In the first stage, the encoder estimates MVs between each frame and its previous N frames, and stores them for possible use in the second stage. Parameter N determines how many previous frames can be used as references for motion-compensated prediction, and should be chosen depending on the maximum expected Round Trip Time (RTT). For example, if the maximum $RTT = 110\text{ms} = 0.11\text{s}$, and the frame rate is $f = 30\text{fps}$, then $N = \lceil RTT \times f \rceil = 4$ frames. The first stage can be thought of as the standard motion compensation loop without entropy coding, as shown in Fig. 2. The reference picture buffer holds N previous reconstructed frames, and MVs are estimated between the current and each of these N previous frames. The output of the first stage are MVs - up to N MV fields for each frame.

The second stage is the standard motion compensation loop without motion estimation, as shown in Fig. 3. Instead of motion estimation, precomputed MVs from the first stage are used in the loop. Depending on the feedback from the decoder, the encoder chooses the appropriate reference frame from the reference picture buffer, and uses the associated (precomputed) MV field to perform motion compensation. *Only this second stage needs to run at streaming time.* By avoiding motion estimation, the second stage of the two-stage encoder has a much lower complexity than the standard encoder, and is more suitable for real-time operation.

Note that the proposed two-stage encoder generates the compressed video bitstream in the second stage, during the streaming session. Prior to streaming, video is stored in raw (uncompressed) form at the server along with the motion vectors generated in the first stage. Hence, reducing the complexity of the second stage, which needs to run in real time during the streaming session, will cost us some extra storage space at the server. To estimate this extra cost, we can use the simplified analyis as follows. Suppose that, on the average, one motion vector is assigned to every 8×8 block of pixels in each frame (in reality, larger or smaller block sizes may be used; some blocks are coded in the I-mode without any associated motion information). If the search range is ± 127, then x- and y-component of a motion vector each take 8 bits (one byte) to store. Hence, we get two extra bytes for each group of 64 pixels, which represents an increase of $2/64 \approx 3\%$ for grayscale video, or $2/96 \approx 2\%$ for YUV 4:2:0 video. So the increase in storage requirements is approximately $2\% - 3\%$ with respect to raw video size. Whether the ability to do fast RPS justifies this increase in size and the fact that the video is not compressed until the streaming begins depends on the particular application and the associated business model. Availability of cheap storage makes this trade-off very attractive.

Another point that needs to be considered is the issue of quantization of reference frames. Note that the standard encoder of Fig. 1 estimates the motion between the reconstructed (i.e., quantized) previous frame, and the original (i.e., unquantized) current frame. Since the quantization parameter QP may vary, one may wonder what kind of effect would this have on the resulting MVs, and subsequently on the compression performance. It turns out that different QP values generally lead to different MVs. In the following section, among other things, we study the effect of QP values used in the first stage on the compression performance achieved in the second stage of our two-stage encoder.

III. Proposed non-causal whole-frame concealment

This section will describe a novel non-causal whole-frame concealment method. To the best knowledge of the author, all previously proposed methods have tried to use the information offered by previously frames to conceal the whole-frame loss, and the proposed method is the first algorithm that uses non-causal information sources. To explain the algorithm more precisely, we can divide it in to two steps. The first step uses an Optical Flow principle [14] and assumes that every object in the video has a constant velocity very similar to what assumed in [13]. Subsequently, the second step uses a boundary-matching algorithm to reestimate the motion vectors of the remaining areas from the first step. sub-section III-A will discuss the first part of the algorithm, and sub-section III-B will discuss the second part of the algorithm.

A. Non-causal optical flow

The first step of the algorithm uses the optical-flow to estimate the motion vectors of the lost frame. It assumes that every object in the video has a constant velocity. It is assumed that the algorithm works sequentially in time forward direction, but it uses non-causal information during error concealment. Therefore, when the algorithm tries to conceal frame t, it has access to all previous frames even their original version or their concealed version.

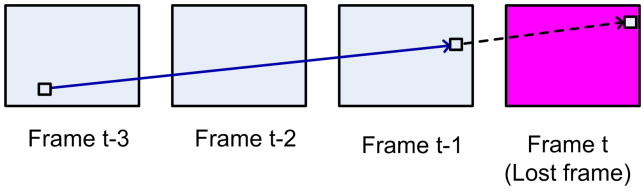
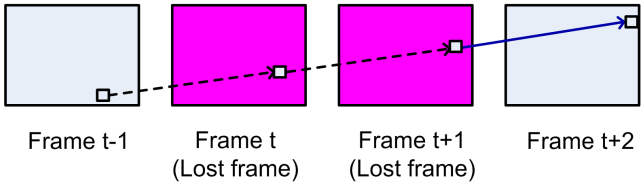


Fig. 4. causal source, original motion vector from frame *t-3* to frame *t-1* showed by solid vector, estimated motion vector from frame *t-1* to frame *t*, showed by dashed vector



[t]

Fig. 5. Anti-causal source, original motion vector from frame *t+1* to frame *t+2* showed by solid vector, estimated motion vectors for two lost frames from frame *t-1* to frame *t* and from frame *t* to frame *t+1*, showed by dashed vectors

If frame *t* has been lost, the algorithm has three different sources of information to estimate the motion vectors.

- **causal source:** Frame *t-1* can act as a causal source of information for motion vector estimation. With the assumption of constant velocity for each pixel in frame *t-1*, the algorithm can find the next place for each pixel in frame *t-1* by extending and scaling its motion vector to frame *t* (except for the intra encoded pixels in frame *t-1*). Estimation from a causal source is illustrated in Fig 4 and Algorithm 1.
- **Anti-causal source:** Here, the algorithm tries to use the next correctly received frame and continues the motion vectors of the next correctly received frame in the time-reversed direction. Estimation from anti-causal source is illustrated in Fig 5 and Algorithm 2.
- **Non-causal source:** If the video communication system uses NACK-RPS as the error resilience technique, the proposed algorithm can detect the reset frame and then use its motion vectors in order to estimate the lost frame's motion vectors. Estimation from a non-causal source is illustrated in Fig 6.

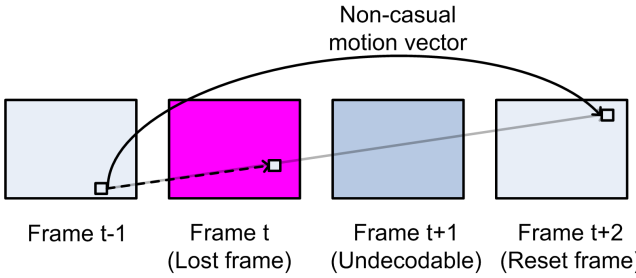


Fig. 6. Non-causal source, original motion vector from frame *t-1* to frame *t-2* showed by solid vector, estimated motion vector from frame *t-1* to frame *t*, showed by dashed vector

After continuing and projecting all motion vectors from a certain source (causal, anti-causal, or non-causal) for the lost frame, for each point in the lost frame, three possible outcomes can result: having only one candidate as its motion vector, having no candidate as its motion vector, or having more than one candidate as its motion vectors. Since occlusion may cause a pixel to have more than one candidate or may cause a pixel to have no candidate (empty area), and intra coded blocks may cause a pixel to have no candidate (empty area), the algorithm uses pixels with only one candidate from a certain source. Now, the question is which source is more accurate and should be used first? Having only one candidate does not guarantee that a point is occlusion free, but the causal source gives the best chance for the algorithm to escape from these uncertain points. Therefore, at first, the algorithm tries to use accurate points (points with only one candidate for their motion vectors) in the causal source. Then, it tries to use accurate points in the anti-causal source to fill the remaining points from the previous step (causal estimation). Finally, it tries to estimate the remaining points from the non-causal

Algorithm 1 causal motion vector estimation

if frame t is lost **then**
  Initialize $\sharp candidate_{causal}(p_t, t) \leftarrow 0$
  for all points p_{t-1} in frame t-1 **do**
    if p belongs to a P-coded block **then**
      $p_t = p_{t-1} + \frac{MV(p_{t-1},t-1)}{index(p_{t-1},t-1)}$
      $MV(p_t, t) = \frac{MV(p_{t-1},t-1)}{index(p_{t-1},t-1)}$
      Increase $\sharp candidate_{causal}(p_t, t)$ by one
    end if
  end for
  for all points p_t in frame t **do**
    if $\sharp candidate_{causal}(p_t, t) == 1$ **then**
      SET ($MV_{causal}(p_t, t)$ is valid)
    end if
  end for
end if

Algorithm 2 Anti-causal motion vector estimation

if frame t is lost **then**
  Initialize $\sharp candidate_{anti-causal}(p_t, t) \leftarrow 0$
  $t_n \leftarrow FindNextCorrectlyReceivedFrame(t)$
  for all points p_{t_n} in frame t_n **do**
    if p belongs to a P-coded block **then**
      $p_t = p_{t_n} + \frac{t_n - t}{index(p_{t_n},t_n)}.MV(p_{t-1}, t-1)$
      $MV(p_t, t) = \frac{MV(p_{t-n},t_n)}{index(p_{t_n},t_n)}$
      Increase $\sharp candidate_{Anti-causal}(p_t, t)$ by one
    end if
  end for
  for all points p_t in frame t **do**
    if $\sharp candidate_{Anti-causal}(p_t, t) == 1$ **then**
      SET ($MV_{Anti-causal}(p_t, t)$ is valid)
    end if
  end for
end if

source because motion vectors in non-causal source have a large temporal length and this fact reduces the accuracy of our assumption (constant velocity of objects).

After performing all the tasks outlined above, the algorithm removes isolated points in the FMV field because they are suspected of being noisy. In our algorithm, an *isolated* point is defined as a point which has less than 5 neighbors in a 5×5 mask neighborhood.

B. Spatiotemporal boundary matching algorithm

The boundary matching algorithm has proven to be a powerful spatiotemporal technique in video error concealment. The Boundary matching algorithm uses two important properties of video signals: temporal correlation and spatial correlation. In video signals, a strong temporal correlation usually exists between current frame and previous frames. Therefore, the empty area in the lost frame after the non-causal motion estimation in sub section III-A can be replaced with an area in a previous frame with the same shape. Spatial correlation between spatially neighboring pixels allows us to use available neighboring pixels, which create the boundary of empty area, in order to find the best match for the empty area from a previous frame [7], [8], [18]. The algorithm uses the sum of squared difference (SSD) of boundaries as the distortion function and tries to find the motion vector corresponding to the minimum measured distortion function [8].

IV. Experimental results

A. 2-stage encoder

Both stages of our two-stage encoder have been implemented using the JM 12.4 reference software. The necessary modifications were made to store MVs on the hard disk in the first stage, and use these MVs instead of motion estimation in the second stage. The performance of the two-stage encoder was compared aginst the standard JM 12.4 encoder on several test sequences - *Bus*, *Football*, and *Foreman* - all in YUV 4:2:0 format, CIF resolution, at 30 fps.

Algorithm 3 Non-causal motion vector estimation

if frame t is lost **then**
  Initialize $\sharp candidate_{non-causal}(p_t, t) \leftarrow 0$
  $t_n \leftarrow FindRPS_{NACK}ResetFrame(t)$
  if $RPS_{NAK}resetframeisavailable$ **then**
    for all points p_{t_n} in frame t_n **do**
      if p belongs to a P-coded block **then**
        $p_t = p_{t_n} + \frac{t_n - t}{index(p_{t_n}, t_n)}.MV(p_{t-1}, t-1)$
        $MV(p_t, t) = \frac{MV(p_{t-n}, t_n)}{index(p_{t_n}, t_n)}$
        Increase $\sharp candidate_{Non-causal}(p_t, t)$ by one
      end if
    end for
    for all points p_t in frame t **do**
      if $\sharp candidate_{Non-causal}(p_t, t) == 1$ **then**
        SET ($MV_{Anti-causal}(p_t, t)$ is valid)
      end if
    end for
  end if
end if

In subsection IV-A1, we validate the performance of the two-stage encoder and investigate the effects of using different QP values in the two stages. Next, in subsection IV-A2 we test the performance of the two-stage encoder with RPS over a lossy network. Finally, in subsection IV-A2 we illustrate the execution speed of the second stage of the two-stage encoder by comparing its execution time to the JM 12.4 encoder.

1) Characterization of the two-stage encoder: To characterize the two-stage encoder, we perform the following tests. For the first stage of the encoder, we turn off the rate control of the encoder, and set the quantization parameter manually to QP $\in \{0,\ 25,\ 30,\ 35,\ 40\}$, where QP $= 0$ corresponds to using the original (unquantized) previous frames for motion estimation and compensation. Using these QP values, we estimate the MVs between each frame and its previous $N = 16$ frames. For each QP value, the output of this stage is the set of MVs, which we store on the hard disk.

In the second stage, we keep the rate control off, and choose various values of QP between 26 and 42 to cover a range of QP values that are typically used in encoding. For each of the QP values in the second stage, we read from the hard disk the MVs corresponding to one QP value from the first stage, and use those MVs for encoding. In this test we are testing coding efficiency rather than lossy transmission performance, so we only use the MVs between the current frame and its most recent previous frame. In Figures 7 - 9, we show the PSNR difference between the the video produced by the JM encoder (which uses the same QP value as the second stage of our encoder), and our two-stage encoder which, in the second stage, uses the QP value indicated on the horizontal axis in the graph. Note that when the QP values from the first stage and the second stage match, the two-stage encoder produces the same video quality (indeed, the same bitstream) as the JM encoder. When there is a mismatch in the two QP values, the performance degrades. Hence, for VBR video (with fixed QP), the two-stage encoder can achieve the same performance as the JM encoder, as long as the same QP value is used in both stages. However, for CBR video, QP values change according to the rate control algorithm employed. In order to achieve the same performance as the JM encoder, we would need to generate and store MVs for all QP values that the JM encoder may use, which would require additional disk space.

Mismatch in the QP values in the two encoder stages may also cause increase in the resulting bit rate. Fig. 10 shows the resulting bit rate of the two-stage encoder, normalized to the bit rate produced by the JM 12.4 encoder for the *Bus* sequence. The data for the other two sequences is similar, and not included in the paper due to space constraints. As in Figures 7 - 9, we see that the two-stage encoder produces the same bitrate as the JM encoder as long as the QP values used in the two stages are the same. Overall, the results show that *the two-stage encoder can have the same compression performance as the standard H.264/AVC encoder* if the second encoding stage has access to the same MVs as the standard encoder. In the CBR case, this means that the first encoding stage needs to produce and store MVs for all QP values that

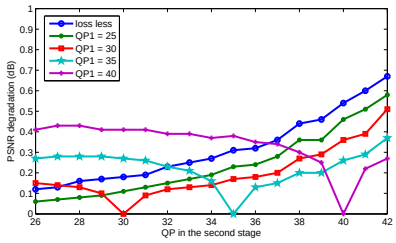


Fig. 7. Quality degradation when QP in the first stage (QP1) is different from the QP in the second stage (*Bus*).

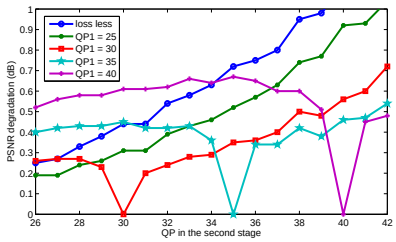


Fig. 8. Quality degradation when QP in the first stage (QP1) is different from the QP in the second stage (*Foreman*).

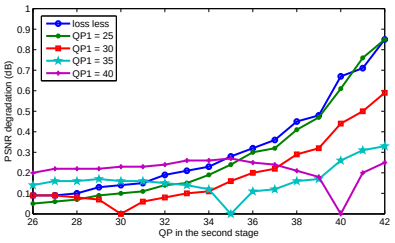


Fig. 9. Quality degradation when QP in the first stage (QP1) is different from the QP in the second stage (*Football*).

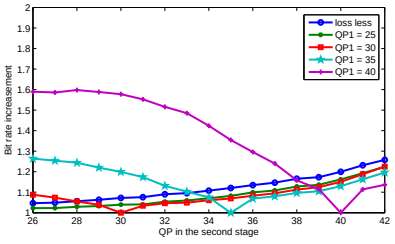


Fig. 10. Bit-rate increase when QP in the first stage (QP1) is different from the QP in the second stage (*Bus*).

are likely to be used by the standard encoder.

2) Performance over a lossy network: To test the performance of the two-stage encoder over a lossy network, we assume that each frame is sent in a single RTP packet. To emulate packet loss, we use packet traces from [19] with average loss rates of $0.05, 0.10$, and 0.20. The server uses ACK-based RPS to decide on the reference frame to be used for motion-compensated prediction. Acknowledgements are assumed to be delivered correctly, but with a variable RTT generated as a normal random variable with a mean of 120 ms and standard deviation of 2 ms. In this test, both the JM encoder and the second stage of our two-stage encoder use rate control, with bit rate $\in \{500,\ 750,\ 1000\}$ kbps. The second stage of the two-stage encoder uses MVs produced by using QP $= 30$ in the first stage. We ran 10 simulation runs for the loss rate of 0.05, and 5 runs for the other two loss rates. The average PSNR results are shown in Fig. 11 for the first 150 frames of *Football*. The performance gap between the JM encoder and our two-stage encoder is no more than 0.3 dB in this lossy transmission scenario with rate control. Similar results were obtained for the other two sequences.

3) Encoding speed: To test the encoding speed, we compare the time it takes to encode the first 250 frames of *Foreman* by the JM 12.4 encoder, and the second stage of our two-stage encoder. Both were compiled using the Microsoft Visual Studio 2005 C++ compiler, and tested on the Intel Centrino machine with 1.70 GHz CPU and 512 MB of RAM. Encoding times are listed in Table I for $N \in \{4, 8, 12, 16\}$ frames in the reference picture buffer. Encoding time for the two-stage encoder includes reading the precomputed MVs from the hard disk. Note that neither encoder is optimized for the particular processor used in this test; in fact, both are based on the general-purpose JM 12.4 software, which is rather slow, so the encoding times listed in the table are somewhat pessimistic. As shown, the second stage of our two-stage encoder achieves approximately a 10-fold speed-up compared to the JM encoder for $N = 4$, and even higher speed-up for larger N.

B. Non-causal whole-frame concealment

There is no experimental result for this part.

V. Conclusion

In this paper, we have described a two-stage H.264 compatible video encoder, that is suitable for real-time video streaming with RPS. The proposed encoder has a compression performance comparable to JM12.4 encoder, while achieving significant encoding speed-up at streaming time. This accerelation is achieved at the expense of additional storage space needed to hold the pre-computed motion vectors and uncompressed video prior to streaming.

In addition, this paper proposes a novel non-causal whole-frame error concealment algorithm. The proposed error concealment algorithm has the ability to extract the information in the preceding frames and succeeding frame and uses this information to conceal the lost frame. NACK-RPS is a successful error resilience method. The proposed error concealment algorithm has the ability to detect use of NACK-RPS and work with NACK-RPS jointly, consequently provides better performance in dealing with errors.

N	JM 12.4 encoder	Two-stage encoder
4	1389	144
8	2500	151
12	1325	157
16	3846	161

TABLE I
ENCODING TIME (IN SECONDS) OF THE FIRST 250 FRAMES OF *Foreman* WITH N FRAMES IN THE REFERENCE PICTURE BUFFER.

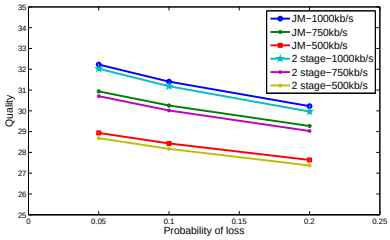


Fig. 11. PSNR over packet traces from [19] using the first 150 frames of *Football*.

REFERENCES

[1] Y. Wang, S. Wenger, J. Wen, and AK Katsaggelos, "Error resilient video coding techniques," *Signal Processing Magazine, IEEE*, vol. 17, no. 4, pp. 61–82, 2000.

[2] S. Fukunaga, T. Nakai, and H. Inoue, "Error resilient video coding by dynamic replacing of referencepictures," *Global Telecommunications Conference, 1996. GLOBECOM'96.'Communications: The Key to Global Prosperity*, vol. 3, 1996.

[3] T. Wiegand, GJ Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, 2003.

[4] B. Girod and N. Farber, "Feedback-based error control for mobile video transmission(is the one used for icip2008?)," *Proceedings of the IEEE*, vol. 87, no. 10, pp. 1707–1723, 1999.

[5] M.M.K. Gong-San Yu Liu and MW Marcellin, "Pocs-based error concealment for packet video using multiframeoverlap information," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 4, pp. 422–434, 1998.

[6] Z. Alkachouh and MG Bellanger, "Fast dct-based spatial domain interpolation of blocks in images," *Image Processing, IEEE Transactions on*, vol. 9, no. 4, pp. 729–732, 2000.

[7] L. Atzori, FGB De Natale, and C. Perra, "A spatio-temporal concealment technique using boundary matchingalgorithm and mesh-based warping (bma-mbw)," *Multimedia, IEEE Transactions on*, vol. 3, no. 3, pp. 326–338, 2001.

[8] Y. Chen, Y. Hu, OC Au, H. Li, and CW Chen, "Video error concealment using spatio-temporal boundary matching and partial differential equation," *Multimedia, IEEE Transactions on*, vol. 10, no. 1, pp. 2–15, 2008.

[9] Y. Wang and Q.F. Zhu, "Error control and concealment for video communication: a review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, 1998.

[10] S. Kumar, L. Xu, M.K. Mandal, and S. Panchanathan, "Error resiliency schemes in h. 264/avc standard," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 425–450, 2006.

[11] WM Lam, AR Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 5, 1993.

[12] Y.K. Wang, MM Hannuksela, V. Varsa, A. Hourunranta, and M. Gabbouj, "The error concealment feature in the h. 264 test model," *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 2, 2002.

[13] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, "Concealment of whole-frame losses for wireless low bit-rate video based on multiframe optical flow estimation," *Multimedia, IEEE Transactions on*, vol. 7, no. 2, pp. 316–329, 2005.

[14] Y. Wang, Y. Zhang, and J. Ostermann, *Video Processing and Communications*, Prentice Hall PTR Upper Saddle River, NJ, USA, 2001.

[15] P. Baccichet, D. Bagni, A. Chimienti, L. Pezzoni, and FS Rovati, "Frame concealment for h. 264/avc decoders," *Consumer Electronics, IEEE Transactions on*, vol. 51, no. 1, pp. 227–233, 2005.

[16] H. Kimata, Y. Yashima, and N. Kobayashi, "Time adaptive motion estimation method for software-based real-time video coding," *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 1151–1154, 2001.

[17] M. Horowitz, A. Joch, F. Kossentini, and A. Hallapuro, "H. 264/avc baseline profile decoder complexity analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 704–716, 2003.

[18] T. Chen, X. Zhang, and Y.Q. Shi, "Error concealment using refined boundary matching algorithm," *Information Technology: Research and Education, 2003. Proceedings. ITRE2003. International Conference on*, pp. 55–59, 2003.

[19] S. Wenger, "Proposed error patterns for internet experiments," *ITU-T Study Group 16 H. 263+ Video Experts Group*, vol. 15, 1999.

S. Mohsen Amiri received his B.Sc. degree from Department of Electronics, Faculty of Computer and Electrical Engineering, Isfahan University of Technology (IUT), Iran, in 2004. He worked as research assistance in IUT, AI-Lab from 2002-2003. He joined IUT Robotic-Center in 2003 and awarded as the 3rd place in Robocop world cup, in Italy 2003. He received his masters in Artificial Intelligence from Department of Computer Engineering, Sharif University of Technology (SUT), Tehran, Iran. And he is currently a master student in Electrical Engineering, Department of Engineering Science, Simon Fraser University (SFU), Burnaby, BC, Canada. His research interests are in Video Communication and streaming, Signal and image processing (primary emphasis on color/multispectral image processing and multidimensional texture segmentation and classification), data mining, algorithm design, and optimization algorithms.

Joint Source-Channel Decoding of JPEG2000 Images with Unequal Loss Protection

Sohail Bahmani

***Abstract*—This paper presents a method for joint decoding of JPEG2000 bitstreams and Reed-Solomon codes in the context of unequal loss protection. When the Reed-Solomon decoder is unable to retrieve the erased source symbols, the proposed joint decoder searches through the set of possible erased source symbols, making use of error resilient features of JPEG2000 to retrieve the correct symbols. To expedite the search process, a relatively small amount of side information is transmitted with high reliability. The joint decoder can be used as an add-on module to some of the existing schemes for unequal loss protection and can improve the PSNR of decoded images by over 10 dB in some cases.**

***Index Terms*—Unequal loss protection, JPEG2000, Reed-Solomon code, joint source-channel decoding.**

I. Introduction

Recent developments in scalable image and video coding, which led to the emergence of scalable coding standards such as JPEG2000 [1] and H.264/SVC [2], have become the motivation for researchers to investigate robust transmission of scalable bitstreams over lossy channels. Due to the progressive traits of the scalable bitstreams, they are more subject to error propagation. The first erroneous or erased bit in these bitstreams can damage all the information carried by the remainder of the bitstream. Therefore, Unequal Error Protection (UEP) and Unequal Loss Protection (ULP) schemes have been developed for protecting scalable bitstreams against errors and erasures, respectively.

In this paper, we have focused on transmission of JPEG2000 images over packet-based networks; therefore, the ULP structure is chosen as the protection framework. To protect scalable bitstreams in ULP schemes with erasure correction codes, several scenarios have already been proposed in [4], [5], and [6]. In these scenarios, knowing the channel's erasure probability distribution, first a quality measure is formulated (e.g., expected distiortion, or expected Peak-Signal-to-Noise-Ratio (PSNR) of the received bitstream.) Then, this measure is optimized as an objective function to find a feasible solution to optimal allocation of protection symbols to different segments of the bitstream. This optimization is usually performed by one of the various optimization algorithms such as local search or hill climbing, under a constraint on the rate budget (i.e., total number of bits allocated for the transmission) [4]-[6].

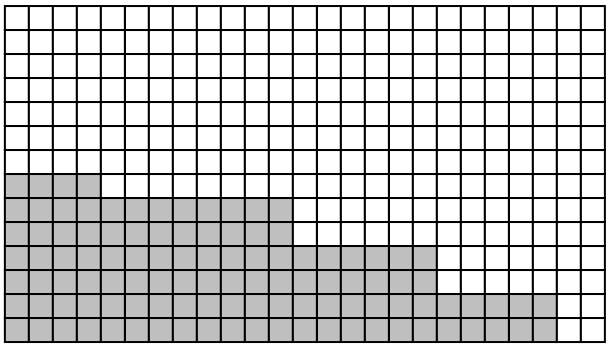


Fig. 1. A typical ULP structure

In our previous work on this subject [3], we presented a joint source-channel decoding (JSCD) technique for worst-case scenarios to improve the quality of received JPEG2000 images transmitted with ULP. By means of the JPEG2000's error resilience (ER) features, this JSCD technique could help to extract source information from the received ULP packets more than plain ULP decoding can achieve. In this paper, we improved the previous JSCD technique such that the time required for the additional decoding is significantly reduced. However, this improvement requires a marginal increase of the rate budget for transmitting some "auxiliary bits." In the proposed scheme, these auxiliary bits are, in fact, the first few significant bits of the source symbols/bytes their erasure would escalate the required decoding time.

This paper is organized as follows: In sections II, the ULP structure is explained in more details. Section III, includes a brief description of JPEG2000 and its ER features. In section IV, the process of JSCD is detailed. Results are presented in section V. Note that the reported results are not outcomes of experiments.

II. Unequal Loss Protection

As Fig. 1 shows, a ULP structure can be represented by a simple table. The packets, which are going to be transmitted through the channel, are represented by the rows of this table. On the other hand, each column of the table is a systematic codeword of an erasure correction code, which contains part of the source symbols and their associated parity symbols.

Since Reed-Solomon (RS) codes are efficient erasure correction codes, they are widely used for ULP.

To achieve the best outcome of using ULP, it is necessary to find the optimal number of redundant symbols in each column (i.e., RS parity symbols). Given the operational distortion-rate curve of the to-be-transmitted image, the objective function of the required optimization can be written as:

$$E[D] = \sum_{i=0}^{N} p_i D\left(\sum_{1 \le j \le L \mid f_j \ge i} \left(N - f_j \right) \right), \qquad (1)$$

where $D(.)$ is the operational distortion-rate function, N is the number of packets, L is the packet's length (i.e., the number of symbols in each packet,) p_i is the probability of having exactly i erasures, and f_j is the number of redundant symbols used in j^{th} column. Since the rate budget for transmission is limited, this optimization is a constrained optimization. One possible constraint is to fix the number of packet's, N , and their length, L . Under these conditions, the optimization algorithm in [6], which uses local search, provides competitive outputs while it requires much less time for the optimization process. Using the JSCD method, the number of received symbols that can be decoded is potentially increased, and consequently, the objective function presented in (1) would not take the improvements into account. Nonetheless, we used the original objective function in [6]'s provided *C* code (i.e., the expected PSNR, which can be formed similar to (1),) because the performance improvement resulting from modification of the objective function is not remarkable.

For $1 \le j \le L$, let f_j^* be the number of parity symbols that is finally assigned to the j^{th} column by the optimization algorithm. If $k_j = N - f_j^*$ represents the number of the source symbols in the j^{th} column, then for each j from 1 to *L*, k_j symbols are read from the JPEG2000's bitstream and encoded by an $RS(N, k_j)$ encoder. The j^{th} column in the ULP frame is then filled by the generated RS codeword and the process continues for the remaining columns.

Suppose that, after transmission via a lossy channel, *e* packets are erased from the completely filled the ULP frame. Therefore, at the decoder side, only column-wise RS codewords with $f_j^* \ge e$ can be successfully decoded. At this point, the task of ULP scheme is over and the recovered a part of the JPEG2000 bitstream is ready to be source decoded. However, after decoding as many columns of the ULP frame as the erasures allow, the ER features of the JPEG2000 bitsream can be exploited to retrieve the transmitted information beyond the received packets' capacity, and thus, will have an image with higher quality after source decoding.

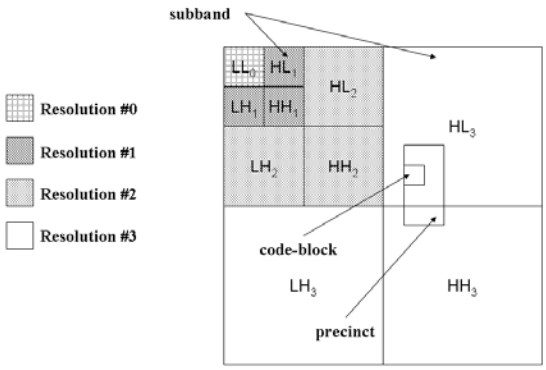


Fig. 2. Typical JPEG2000's tile-components structure

III. Error Resilience In JPEG2000

JPEG2000 is an image compression standard mainly developed based on discrete wavelet transform (DWT.) In this standard, the raw image is first cut off into rectangular *tiles* to be compressed separately, though multiple tiles are generally used for sufficiently large images. After the necessary colour transformations are applied to each tile, the resulted *tile-components* are transformed by a wavelet/subband transform to form the *subbands*. As illustrated in Fig. 2, in the JPEG2000's hierarchical structure, within the tile-components, each set of subbands that have a common frequency band are referred to as a *resolution*. Subbands are further partitioned into square-shaped segments named *code-blocks*, which are the smallest divisions in the JPEG2000 standard encoded independently. Also, the JPEG2000's encoder can produce bitstreams with various scalability properties, each with different ordering of coding-passes in the bitstream. Therefore, to conform to our requirements, *resolution-scalable* JPEG2000 bitstreams are used to prevent coding-passes of a particular code-block being scattered throughout the bitstream. Furthermore, an optional rectangular subdivision of resolutions, called a *precinct*, includes a whole number of code-blocks. Although the precincts enable some of the JPEG2000's extra capabilities, including the error resilience (ER) feature, they do not have a direct role in the compression process.

The basis for ER features of JPEG2000 is established upon the fact that entropy-coded bitstreams are sensitive to errors (i.e., a corrupted entropy-coded bitstream could trigger a failure during its decoding process.) The quantized wavelet samples of each code-block are encoded by MQ-coder, a context-based arithmetic coder, bit-plane by bit-plane which is the crucial factor in scalability of the JPEG2000 bitstreams. This bit-plane coding operation, in JPEG2000, is carried out in several alternating passes of three different types. When an error occurs in the bitstream generated in one of these so-called *coding-passes*, the decoder is able to locate the error within the next few following coding-passes [7], [8]. Moreover, on decoding failures, the remaining section of the code-block's bitstream would be partially decoded or

completely discarded depending on the combination of applied ER modes. These ER modes, which together form the ER features of JPEG2000 bitstreams, are thoroughly explored in [8].

In this paper, two ER modes, namely ERTERM and RESTART, are used since they do not interfere with the constraints of our JSCD technique. With these two modes switched on, inter-code-block error propagation is prevented. Activated RESART mode, forces the MQ-coder to restart at the beginning of each coding-pass; therefore, each coding-pass obtains a separate codeword segment. In addition, the ERTERM switch decides on the predictable termination policies for each MQ and/or raw codeword segment. These policies are the means to detect errors in MQ codeword segments with high probability.

IV. Joint Source-Channel Coding For ULP

The ULP scheme, on average, can deliver an adequate number of source bits for a sufficiently clear image. Nonetheless, in the worst cases, when excessive packet erasures occur, a significant piece of the transmitted source bits would become corrupted. Consequently, the output image's visual quality would be degraded substantially. This degradation can be alleviated by our proposed JSCD method.

In the JSCD technique presented by [7], the ER features of JPEG2000 bitstreams are utilized to accelerate convergence of the LDPC soft decoder in an iterative fashion. Analogously, in JSCD for ULP, JPEG2000's ER features are exploited to verify the validity of the guesses, which are made to decode more ULP columns.

As mentioned in section I, each of the ULP frame's columns is an RS codeword. Therefore, they can be decoded as long as the number of their parity symbols is not less than the number of erased symbols (i.e., the number of erased packets.) Suppose that the ULP frame includes an RS (N,k) codeword as one of its columns. If $e = N - k + d$ packets are erased during the transmission ($d \geq 1$), then by filling the locations of d erased symbols in the taken column with arbitrary values, the RS decoder can fill the remaining erasure locations. However, the resulting complete RS codeword is, perhaps, not the same as the transmitted RS codeword. Here, correctness of the guessed symbols is verified by the ER features of JPEG2000; if the JPEG2000's decoder complains of an error, then the next possible unverified RS codeword is generated and examined by the JPEG2000's decoder.

To follow this procedure more carefully, first take f_j^*'s as the solution to the optimization problem discussed in section II. Let $f^{CS} = \max_{e > f_i^*} f_i^*$ be the number of parity symbols in those columns with protection levels just less than e, the number of erasure. The superscript *CS* stands for *critical segment,* a part of the ULP frame that consists of columns that have f^{CS} parity symbols. To retrieve the information contents of a critical segment, it is sufficient to find the correct value of $d = e - f^{CS}$ erased symbols.

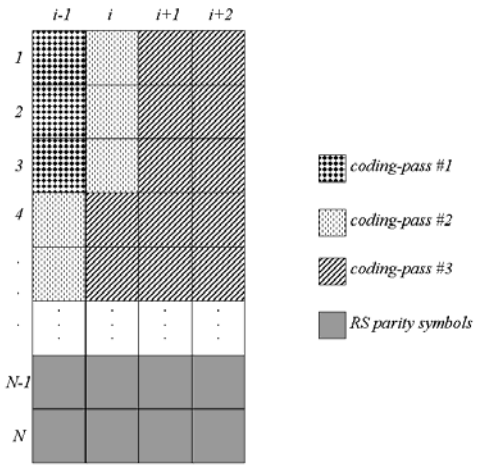


Fig. 3. Typical coding-pass segments inside a section of ULP structure

Filling locations of d erased symbols would leave only f^{CS} erased symbols in the critical segment's columns, which enables the RS decoder to fill in the remaining erased symbols. However, there is no guarantee that the resulted columns, which are filled by guessing d erased symbols, coincide with what has been transmitted. To ensure the correctness of a repaired column of the critical segment, we let JPEG2000 attempt to decode the source information contained in that column. Since the transmitted JPEG2000 bitstream is supposed to be error resilient, an invalid segment in the bitsream would cause the decoder's complaint; in that case, the incorrect guess is discarded and another unexamined guess among all $2^{m \times d}$ possible guesses is selected for the next iteration. Since RS(N,k) codes are able to reproduce any set of $N-k$ erased symbols, choosing any arbitrary values for the d erased symbols would result in a valid RS codeword, though only one of them is correct from JPEG2000's perspective.

The number of possible guesses is a crucial factor which determines the time consumption of the JSCD algorithm. However, as concluded in the previous paragraph, the information carried by non-erased symbols is not advantageous for reducing the number of guesses. Therefore, some side information is necessary to speed up the process of guessing the erased symbols. This extra information can be generated using different strategies; however, the strategy that we adopted is to select those symbols that, possibly, have more contribution to the total number of guessing iterations, and transmit their first few significant bits with high protection. Depending on the number of these auxiliary bits spent for the selected symbols, and the layout of coding-passes inside columns of the ULP frame, the JSCD time could decline exponentially; but the ratio of the number of spent auxiliary bits to the total rate budget should not be ignored.

In Fig. 3, a part of a ULP structure is illustrated in which a particular coding-pass (i.e., the hatched cells), has occupied more than one symbol in some of the rows/packets. Suppose that one of these packets and, consequently, several symbols

TABLE I. Average PSNR changes resulted by JSCD with ($\overline{\Delta PSNR_2}$ (dB)) and without ($\overline{\Delta PSNR_1}$ (dB)) auxiliary bits, and width for the last three segments and percentage of required increase in rate budget for auxiliary bits ($\Delta R / R$.)

Image	Loss rate	*CS #1*			*CS #2*			*CS #3*			$\Delta R / R$ *(%)*
		$\overline{\Delta PSNR_1}$	$\overline{\Delta PSNR_2}$	*Width*	$\overline{\Delta PSNR_1}$	$\overline{\Delta PSNR_2}$	*Width*	$\overline{\Delta PSNR_1}$	$\overline{\Delta PSNR_2}$	*Width*	
LENA	10%	+10.0	+10.1	91	+0.7	+0.7	4	0.0	0.0	1	0.65
	20%	+6.4	+6.5	70	+1.2	+1.2	10	+0.1	+0.1	3	0.70
	30%	+4.2	+4.6	60	+0.5	+0.9	5	+0.2	+0.4	4	0.72
BARBARA	10%	+7.4	+7.4	93	-0.1	0.0	1	+0.3	+0.3	2	0.78
	20%	+5.8	+5.8	73	+0.2	+0.3	3	+0.7	+1.0	2	0.84
	30%	+3.5	+3.5	80	+0.2	+0.7	8	0.0	0.0	4	0.95
GOLD HILL	*10%*	+3.7	+3.7	75	+0.1	+0.2	9	+0.2	+0.2	6	0.79
	20%	+0.0	+0.2	2	+0.6	+0.6	14	+0.4	+0.4	55	0.82
	30%	-0.5	0.0	1	-0.4	-0.2	2	+0.9	+.9	63	0.89
BOAT	10%	+9.8	+9.8	85	+0.4	+0.6	4	-0.5	-0.1	1	0.76
	20%	0.0	+0.1	2	+6.7	+6.7	70	+1.2	+1.6	10	0.86
	30%	+8.7	+8.8	65	+2.1	+2.1	20	+0.4	+0.4	2	0.92

of the mentioned coding-pass are erased. If the erased symbols had to be guessed, all of them should be filled with correct values to decode the coding-pass perfectly. Therefore, it can be expected that in these erasure incidents, the coding-pass's decoding time raises exponentially. To diminish this effect, in our proposed strategy, the auxiliary bits are assigned to the symbols of each packet, which belong to the same coding-pass. For example, in Fig. 3, the symbols in columns *i* through *i+2* of the fourth row would receive the auxiliary bits, but the symbol in the i^{th} column and the third row would not. As a result, this policy provides a compromise between improving decoding speed and the number of extra bits required.

I. RESULTS

To test our JSCD algorithm, first, we encoded the standard 512×512 gray-scale *Lena, Barbara, Gold Hill, and Boat* images with the Kakadu implementation of JPEG2000 [1] at a rate of 0.7782 bits per pixel (bpp.) This rate is chosen, because we used 255 packets of 100-bytes each in the ULP scheme, which imposes the upper bound for the transmitted images' rate. Moreover, we used the image compressor with 128×128 precincts and activated ERTERM, RESTART, SOP and EPH modes. After that, we derived operational PSNR-rate curves of the produced resolution-scalable JPEG2000 images. Then, using a C implementation of the algorithm proposed in [6], the optimized protection levels of the ULP structure were generated for each image at 10%, 20%, and 30% mean loss rates.

In all of our implementations, although other symbol lengths are not impractical, for convenient adaptation with Kakadu, 8-bit symbols (i.e., bytes), were used as the ULP frame's symbols, or equivalently, the RS codeword's symbols, which are encoded and/or decoded by Phil Karn's RS codec [9]. Also, to make the whole decoding time more attainable, only one byte is guessed in each column of critical-segments (i.e., $d = 1$). In addition, the joint source-channel decoder is only allowed to correct the undetected wrong guesses of the last two columns, when all possible guesses for the present column are made unsuccessfully. For the extended JSCD method, we considered two auxiliary bits for the symbols which need it. Each of the reported results are based on averaging outcomes of 100 simulations that were run on a desktop PC with an Intel Core 2 Duo 2.13 GHz CPU with 2GB of RAM. Nonetheless, Fig.'s 4 and 5 are included to demonstrate the visual improvements by examples.

Table I shows the average change in the PSNR, when JSCD is performed without use of auxiliary bits ($\overline{\Delta PSNR_1}$) or with help of auxiliary bits ($\overline{\Delta PSNR_2}$), both in dB, for different tested images at different mean channel loss rates. The experiments carried out for the last three possible critical segments, are represented in the table by CS#1, CS#2 and CS#3. Each critical segment's width is also reported in the table. Additionally, a column is assigned to the percentage of required increase in rate budget for sending auxiliary bits ($\Delta R / R$). For sufficiently wide critical segments, both methods demonstrate remarkable improvement in the PSNR; however, their difference is negligible in most cases. On the other hand, for narrow critical segments (e.g., those with one or two columns), both methods show a small improvement or even a negative change in the PSNR. The quality reduction is due to imperfections in the JPEG2000's ER capabilities, which causes some incorrect guesses to remain undetected when the short critical segment's decoding is finished.

In Table II, the average time cost of both JSCD variations are reported along with the time that original ULP decoding requires, all in seconds. The time cost of ULP decoding without JSCD, and the mean extra times required by JSCD with and without the use of auxiliary bits are reported under columns T, $\overline{\Delta T_1}$ and $\overline{\Delta T_2}$.

Results show that the auxiliary bits can reduce the time required for the JSCD by up to 13 times, at the cost of less than only 1% increase in the assigned bit budget; however, the

Fig. 4. Barbara image transmitted through a channel with 10% loss rate, before JSCD on left (PSNR = 22dB) and after JSCD on right (PSNR = 28.9dB)

Fig. 5. Boat image transmitted through a channel with 20% loss rate, before JSCD on left (PSNR = 24.3dB) and after JSCD on right (PSNR = 30.1dB)

TABLE II. Average extra time required for JSCD with ($\overline{\Delta T_2}$ (s)) and without ($\overline{\Delta T_1}$ (s)) auxiliary bits, and plain ULP decoding time (T(s))

Image	Loss rate	*CS* #1			*CS* #2			*CS* #3		
		T	$\overline{\Delta T_1}$	$\overline{\Delta T_2}$	T	$\overline{\Delta T_1}$	$\overline{\Delta T_2}$	T	$\overline{\Delta T_1}$	$\overline{\Delta T_2}$
LENA	10%	0.16	108.45	22.64	0.13	9.87	1.63	0.06	0.17	0.17
	20%	0.18	164.71	38.47	0.13	23.29	4.81	0.06	0.52	0.51
	30%	0.17	170.89	13.24	0.09	9.34	1.87	0.07	0.55	0.55
BARBARA	10%	0.07	143.55	21.27	0.09	0.27	0.26	0.06	0.43	0.43
	20%	0.09	239.78	45.72	0.10	0.82	0.82	0.19	0.72	0.72
	30%	0.18	508.36	80.02	0.12	108.80	11.77	0.08	1.39	1.39
GOLD HILL	10%	0.21	192.48	26.94	0.18	62.61	5.79	0.10	1.15	1.15
	20%	0.22	0.51	0.51	0.16	213.38	17.40	0.14	1657.92	119.00
	30%	0.13	0.08	0.08	0.17	1.27	1.26	0.13	1098.41	113.21
BOAT	10%	0.10	203.81	36.91	0.08	4.61	0.74	0.16	0.15	0.15
	20%	0.12	0.07	0.07	0.09	157.65	32.49	0.17	1.5	0.43
	30%	0.17	170.68	43.19	0.08	47.39	8.79	0.17	0.19	0.19

JSCD still takes much more time than the regular ULP decoding. Note that all the executables were compiled and run in "Debug" mode, which is not optimized for speed. Therefore, the reported times are only a measure for comparison, and should not be considered as the best achievable times.

REFERENCES

[1] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, Kluwer, Norwell, MA, 2001.

[2] H.-C. Huang, W.-H. Peng, T. Chiang, and H.-M. Hang, "Advances in the scalable amendment of H.264/AVC," *IEEE Commun. Mag.*, vol. 45, no. 1, pp. 68-76, Jan. 2007.

[3] S. Bahmani, I. V. Bajic, and A. HajShirmohammadi, "*Joint source-channel decoding of JPEG2000 images with unequal loss protection", IEEE ICASSP'08,* Las Vegas, NV, March 2008.

[4] R. Puri and K. Ramachandran, "Multipe description source coding using forward error correction codes," *Proc. Asilomar*, vol. 1, pp. 342-346, Oct. 1999.

[5] A. E. Mohr, E. A. Riskin, and R. E. Ladner, "Unequal loss protection: Graceful degradation of image quality over packet erasure channels through forward error correction," *IEEE Trans. Select. Areas Commun.*, vol. 18, no. 6, pp. 819-828, Jun. 2000.

[6] V. M. Stanković, R. Hamzaoui, and Z. Xiong, "Real-time error protection of embedded codes for packet erasure and fading channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 8, pp. 1064-1072, Aug. 2004.

[7] L. Pu, Z. Wu, A. Bilgin, M. W. Marcellin, and B. Vasic, "LDPC-based iterative joint source-channel decoding for JPEG2000," *IEEE Trans. Image Processing*, vol. 16, no. 2, pp. 577-581, Feb. 2007.

[8] Z. Wu, A. Biglin, and M. W. Marcellin, "Error resilient decoding of JPEG2000" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1752-1757, Dec. 2007.

[9] P. Karn, *C++ class library for Galois field arithmetic and algebra, with RS encoder/decoder*, [Online] Available: http://www.ka9q.net/code/fec/fec-3.0.1.tar.bz2

Sohail Bahmani is a Master of Applied Science student in the school of engineering science at Simon Fraser University, Burnaby since 2007. He received his BS degree in Electrical Engineering from Sharif University of Technology, Iran in 2006. His main research area is joint source channel coding/decoding of scalable multimedia bitstreams.

Predictive Decoding for Delay Reduction in Video Communication

Yue-Meng Chen

***Abstract*—Low delay is critically important for interactive video communication. This paper presents several predictive decoding techniques for end-to-end delay reduction. Video frames are predicted from past video data, and displayed before they arrive at the decoder. This problem is similar to the problem of whole-frame concealment, and we compare our methods to the state-of-the-art whole-frame concealment algorithms. As demonstrated in the paper, our methods yield better results because of incorporated motion segmentation and region-based motion prediction, with gains up to 2 dB compared to the existing whole-frame concealment algorithms on sequences with complex and high-intensity motion. Our simulations indicate that through predictive decoding it may be possible to reduce the perceived end-to-end communication delay by about 100 ms while maintaining reasonable video quality.**

***Index Terms*—Motion segmentation, frame prediction, temporal prediction, end-to-end delay, vector median filter**

I. Introduction

Interactive video applications, such as video phone and video conference, have become a feasible alternative to traditional telephony services. However, providing high-quality interactive visual communications is still a challenging task due to the impairments introduced by the network. In particular, end-to-end delays beyond 150 ms worsen the user experience [1]; delays above 400 ms make interactive communication virtually impossible [2]. The overall end-to-end delay has several components, including acquisition, encoding, transmission delay (which includes propagation delay plus any processing within the network, such as routing), decoding, and rendering. Video acquisition and encoding typically take longer than audio acquisition and encoding, so in practice, audio is often purposefully delayed in order to maintain "lip synchronization" [1]. Hence, reducing the delay associated with video would be beneficial for the entire video communication system, because it would reduce the need to delay the associated audio information.

On the Internet, transmission delay alone may approach 150 ms, as indicated in Table I. The table shows several Round Trip Times (RTTs) measured on the afternoon of Feb. 24, 2007, by pinging various hosts from the Blenz Café in Yaletown (Vancouver, BC), using the FatPort broadband wireless access network. In this scenario, one can expect about a 150 ms one-way transmission delay between Vancouver and Christchurch, New Zealand. The distance between Vancouver and Christchurch, across the Earth's surface, is about 12000 kilometers. If we were to lay down a direct optical cable between these two cities and eliminate all processing, we would get a one-way delay of about $12000/(2\times10^8) = 60$ ms. The difference between this (lowest possible) delay and the measured delay is indicative of the network topology and the processing inside the network.

TABLE I - Several ping RTTs from Vancouver

Host	**RTT** (ms)
ucla.edu	63
mit.edu	113
bbc.co.uk	187
epfl.ch	231
monash.edu.au	280
canterbury.ac.nz	304

In this paper, we present several frame prediction techniques that can help reduce the perceived end-to-end video delay. Using the received video data, future frames are predicted and displayed before they arrive at the decoder, as illustrated in Fig. 1. We refer to this process as ***predictive decoding***. This process inevitably reduces the quality of the displayed frames, especially when the motion is complex, but it also provides the user with the ability to trade-off quality for delay. Our results indicate that using these methods, it is possible to reduce the perceived end-to-end video delay by about 100 ms while maintaining reasonable video quality.

A future frame, which has not been received yet, may be thought of as a frame subject to 100% loss. In this sense, the goals of predictive decoding and whole-frame concealment [5]-[10] are the same. Differences arise in the constraints placed on the algorithms - for example, predictive decoding has to use strictly causal processing, while whole-whole frame concealment does not (although, in fact, most existing whole-frame concealment algorithms rely chiefly on causal processing). We will therefore compare our proposed predictive decoding methods to a state-of-the-art whole-frame concealment method, and demonstrate substantial improvements, especially on sequences which contain complex and high-intensity motion.

The author is with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. E-mail: yuemengc@sfu.ca. Part of this work has been presented at IEEE Globecom 2007.

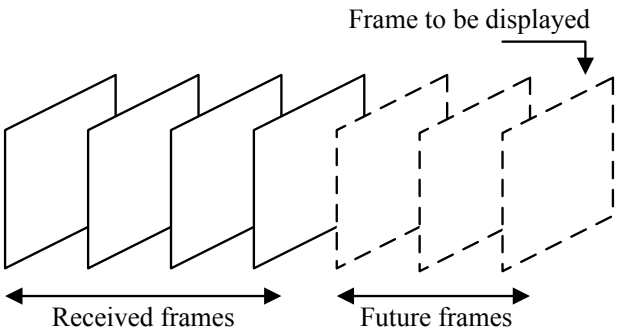


Figure 1: Perceived delay reduction by predictive decoding.

Any frame prediction technique needs to deal with two fundamental challenges: (1) predicting future motion, and (2) synthesizing a frame based on the predicted future motion.

Challenge 1 (Motion prediction): Predicting motion trajectories of individual objects in the scene plays a crucial role in synthesizing future frames. Standard encoders (MPEG-2/4, H.263, H264) estimate motion vectors by minimizing motion compensation errors [3], [4]. The frame is divided into small blocks, and a motion vector is assigned to each block (except for I-blocks). However, boundaries between the objects in the scene hardly ever match the boundaries between blocks, which makes it difficult to separate one moving object from another. In our proposed algorithm for frame prediction, we use motion segmentation to address this problem. This motion segmentation step is also the chief difference between our method and the existing methods for whole-frame concealment. Motion segmentation helps avoid the contamination of object's predicted motion trajectory by another object's motion. As a result, predicted frames exhibit blocking near object boundaries. In addition, motion segmentation allows us to direct the computational effort where it is most needed - near object boundaries – by using denser motion field representation there, and a coarser representation in the object's interior.

Challenge 2 (Future frame synthesis): Using the predicted motion, a future frame can be synthesized by moving the objects from the last received frame along the predicted trajectories onto the future frame.. In this way, two kinds of spatial discontinuities might emerge:

- When two or more objects land on the same part of the frame, a decision needs to be made about the pixel values to be written into this area of the frame. We call these areas *overlapped areas*, and describe how we process them in Section III-4.
- Complementary to the case above, there are also areas of the frame that remain empty after the object move along their predicted trajectories. We call these areas *empty areas*, and describe how we process them in Section III-5.

This paper is an extension of our previous work in [11]. The proposed predictive decoding has been implemented in the popular XviD MPEG-4 video codec [12] and tested on a variety of sequences, ranging from low to high motion. The proposed prediction algorithm has a flexible application interface, and it can be readily adapted to other block based video coders, like H.263+ and H.264.

The material in the rest of this paper is organized as follows: In Section II we briefly review the previous work on whole-frame error concealment. In Section III, we describe the proposed object based frame prediction algorithm. In Section IV, we present the experimental results and give the performance comparison between our frame prediction and a state-of-the-art whole-frame concealment algorithm. The conclusions are drawn in Section V.

II. Previous Work on Whole-Frame Concealment

Strictly speaking, delay reduction through frame prediction does not seem to have been explored in the literature thus far. Yet, several algorithms have been proposed for whole-frame concealment recently [5 - 10], and these algorithms can be re-configured to perform frame prediction as a strategy for reducing the end-to-end communication delay. This observation motivates us to compare our work against the previously proposed methods for whole-frame concealment, which can be broadly divided into two categories: pixel-based and block-based. Below we review a representative method from each category.

Algorithm-A. Pixel based Whole-frame Concealment

Based on optical flow theory, a whole-frame concealment algorithm is proposed for video streaming applications in [5 - 6]. Operating in the pixel domain, a constant velocity motion model is adopted to project the last correctly received frame onto a future frame by the following steps:

Step-1 Constructing a forward motion vector (FMV) for each pixel in the last received frame (i.e., frame $t-1$) from the backward MVs and coding modes of frames from $t-2$ to $t-L-1$ with the constant velocity model,

Step-2 Regularizing and smoothing the FMV field spatially of frame $t-1$ using the two-dimensional median filter,

Step-3 Reconstructing the missing frame by projecting pixels from frame $t-1$ into the same pixels in the missing frame t in a 2x2 pattern, and averaging pixels if two pixels have landed on the same position (overlapped areas),

Step-4 Scanning the missing pixels in frame t and estimating these missing pixels using the median of the neighboring available pixels (empty areas),

Step-5 Filtering and downsampling the reconstructed frame by two in both the horizontal and vertical directions.

One obvious drawback of a pixel-based whole-frame concealment is its computational complexity. Applying a computationally-intensive algorithm to reduce the end-to-end delay is likely to be counter-productive, because it would place additional burden on the decoder performance during the real-time operation. Further, averaging pixels in overlapped areas and median filtering of pixels in empty areas may be able to mitigate to some extent the problems of empty and overlapped areas, but may also lead to objectionable artifacts in the parts of the frame with complex motion or

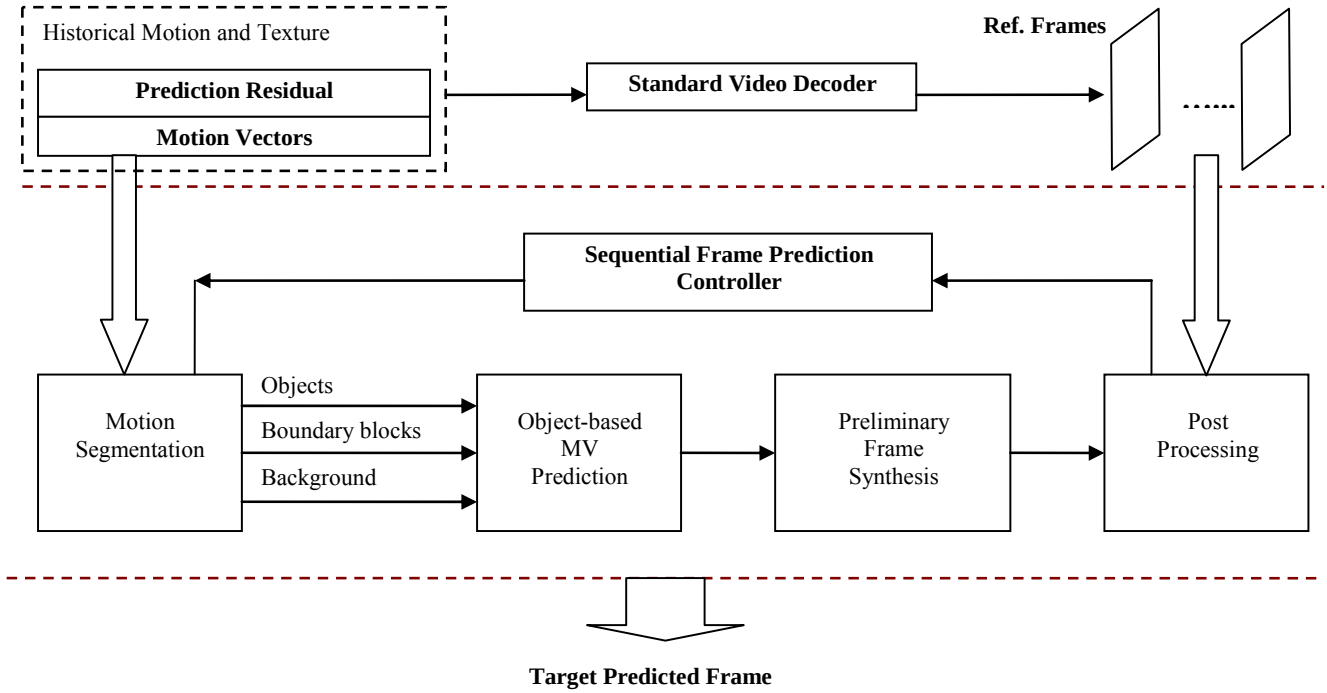


Figure 2: Flow diagram of prediction module

texture.

Algorithm-B. Block based Whole-frame Concealment

A block-based frame concealment was proposed by Baccichet *et. al*. [7 – 8] based on the same optical flow concept, but offering lower complexity with a potential for real-time implementation. This algorithm consists of the following steps:

Step-1 Determining the basis reference frame that presents a significant number of inter coded blocks (e.g., frame t – 1), and constructing a FMV for each pixel in frame t – 1,

Step-2 Projecting regularized pixel-wise FMV from frame t – 1 into pixels in the missing frame t,

Step-3 Computing the statistics (mean and variance of projected MVs) for each 16×16 MB and each 4×4 block in frame t,

Step-4 Recovering MVs for 16×16 macroblocks and 4×4 blocks in frame t based on the statistics information,

Step-5 Reconstructing the final picture using motion compensation and loop filtering in H.264/AVC.

Compared to **Algorithm-A**, **Algorithm-B** reduces the computational complexity greatly. However, by averaging neighboring motion vectors, motion vector recovery in **Algorithm-B** may disrupt the spatio-temporal relationships among the motion vectors of neighboring blocks, and thus may worsen the blocking artifacts. Moreover, the algorithm requires extra effort on the encoder side to produce better motion vectors, and one additional step to estimate motion vectors for intra-coded blocks at the decoder [9].

Since our work is targeted at real-time video communication, and our proposed algorithm is itself block-based, we compared our algorithm against **Algorithm-B**. To this end, we have implemented **Algorithm-B** in the XviD MPEG-4 codec [13], the same platform we used for testing our own algorithm.

III. Proposed Algorithm

The approach we propose, *object-based frame prediction* (OBFP), is shown schematically in Fig. 2. The OBFP system consists of four main processing units controlled by a sequential frame prediction controller. First, the motion segmentation isolates all moving objects from the background on a block-by-block basis in frame t – 1 and identifies blocks that lie near the object boundaries on texture residual energy. Second, a variable-block-size MV prediction estimates the motion trajectory for blocks in frame t – 1. Third, a preliminary future frame is synthesized by moving all blocks from frame t – 1 to frame t along the estimated motion vector. Finally, the post processing unit mitigates the spatial discontinuities (empty and overlapped areas) of the preliminary future frame.

A crucial component of frame prediction is the accuracy of the predicted motion trajectories. By employing an object-based MV prediction in the proposed algorithm, we are able to reduce the effects of motion vector noise and unnatural motion artifacts in the predicted frame, such as image background shaking and spatial discontinuities. Motion trajectory prediction based on moving objects offers a way to enforce motion field homogeneity within objects, and may compensate for MV inaccuracy caused by the standard motion estimation criteria of minimum prediction error at the encoder.

After generating a preliminary version of the future frame by moving the blocks of the last received frame along the

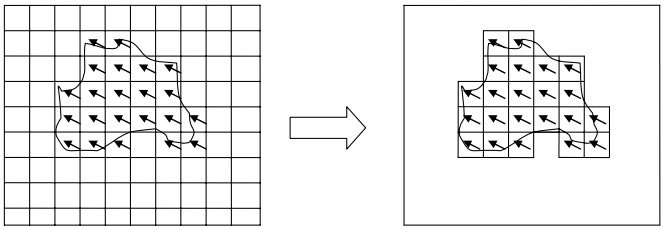


Figure 3: Block-based motion segmentation.

predicted motion trajectories, we apply several post processing steps to deal with empty and overlapped areas. In particular, linear interpolation and motion re-estimation via boundary matching are employed to process empty areas, while pixel averaging and the block selection are combined to process overlapped areas. In the following paragraphs we outline the details of our method.

1) Motion Segmentation

Motion information is available in the compressed video bit stream in the form of motion vectors. By isolating the regions of consistent motion, we can determine where the moving objects are in the frame. Fig. 3 shows an example of segmenting a moving object out of a frame based on motion consistency. Note that, since MVs are block-based, object boundary will also follow block boundaries

References [16] and [17] have described methods for achieving robust motion segmentation with adaptive k-means clustering. For the purpose of identifying moving objects, we develop a block-based motion segmentation method as a combination of K-means clustering [18] and motion consistency verification [19]. The flow diagram of our motion segmentation is shown in Fig. 4.

Motion information associated with each block (MV for P-blocks, zero MV for skip-blocks, no MV for I-blocks) is first extracted from the encoded bit stream. The original MVs are mapped to the minimum block size supported by the particular coder, for example 8×8 blocks for H263 and MPEG4, and 4×4 blocks for H264. This is done simply by splitting the larger blocks to the smallest supported blocks and assigning to each of them the MV of the parent block. The neighboring MVs are then grouped into clusters iteratively using the k-means clustering algorithm with the help of the motion consistency model. The segmentation is carried out as the follows:

Step-1 The motion vector statistical information are computed first to estimate the overall motion level in the clustering threshold estimator

$$\mu_{mv} = E\left\{\overrightarrow{MV}_i\right\} = \frac{1}{N}\left(\sum_{i=1}^{N}\overrightarrow{MV}_i\right) \quad \textbf{(1)}$$

$$\sigma_{mv}^2 = \frac{1}{N}\sum_{i=1}^{N}\left\|\overrightarrow{MV}_i - \mu_{mv}\right\|^2 \quad \textbf{(2)}$$

Where N denotes the number of inter coded blocks in the frame to be segmented. Based on the variance of MVs, the corresponding region growing step size is determined as the value of the standard deviation σ_{mv}, and its maximum value

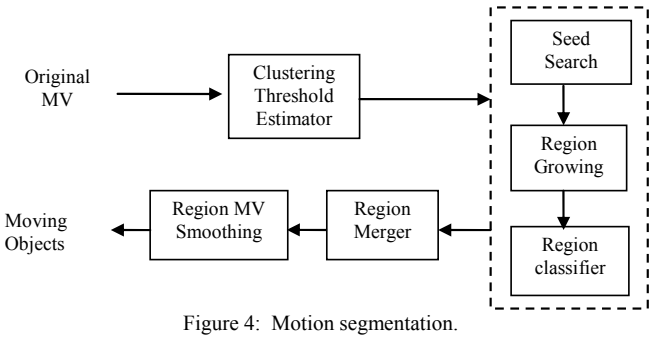


Figure 4: Motion segmentation.

is limited to 8. The threshold of minimum moving object distance is also determined as the same value of σ_{mv}, and limited to a maximum value of 4. These two parameters are used in the region growing unit and the region merger unit, respectively.

Step-2 With the seed pattern [19] shown in Fig. 5, a group of P-blocks with minimum MV distortion is found as the starting point for region growing. The moving objects are identified by (3).

$$D_{seed} = \arg\min_i\left(\sum_{j=1}^{N}\left\|\frac{1}{N}\left(\sum_{j=1}^{N}\overrightarrow{MV}_{i,j}\right) - \overrightarrow{MV}_{i,j}\right\|\right), \quad N = 5 \quad \textbf{(3)}$$

Where D_{seed} is denoted as the overall MV distortion of the seed, and $\overrightarrow{MV}_{i,j}$ denotes the MVs of blocks in a seed candidate (Right, Left, Centre, Upper, Bottom). We go through the entire frame to find a seed with minimum MV distortion. After we identify a new seed, neighboring blocks in horizontal and vertical directions, shown in Fig.6, are first tested for clustering into this new moving region.

Step-3 With the threshold of MV distortion range from step 1) and the seed from step 2), a coherent motion region is grown gradually by clustering bordering blocks into the region if their motion is sufficiently similar to the prevalent motion inside the region. A motion consistency model is established based on the minimum distortion criterion during this region growing procedure. We start with calculating the region growing threshold.

$$D_{TH} = E\left\{\left\|\overrightarrow{MV}_{\text{int}} - \overrightarrow{MV}_{centroid}\right\|\right\} + D_{off} \quad \textbf{(4)}$$

$$\overrightarrow{MV}_{centroid} = E\left\{\overrightarrow{MV}_{\text{int},i}\right\} = \frac{1}{M}\sum_{i=1}^{M}\overrightarrow{MV}_{\text{int},i} \quad \textbf{(5)}$$

Where $\overrightarrow{MV}_{centroid}$ and $\overrightarrow{MV}_{\text{int}}$ are, respectively, the centroid MV of the region and MVs of all internal blocks within the region, and D_{off} indicates the region growing step size that is estimated in the Clustering Threshold Estimator in the step 1. Once a new seed-block is found, the centroid MV, $\overrightarrow{MV}_{centroid}$, is initialized, and the region growing threshold D_{TH} is calculated as in (4). D_{TH} indicates the average distance between the MVs in the region and the centroid MV of the region, increased by the offset D_{off}.

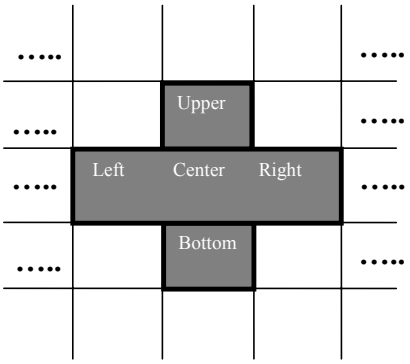


Figure 5: The seed pattern.

The region growing process checks whether an ungrouped block which borders the existing region has sufficiently similar motion to the prevalent motion inside the region, by checking if the following condition is satisfied.

$$D_{i,n} = \left\| \overrightarrow{MV}_{centroid,n} - \overrightarrow{MV}_{ext} \right\| \le D_{TH,n} \qquad \textbf{(6)}$$

Where $\overrightarrow{MV}_{centroid,n}$ denotes the motion vector centroid of region *n*, $\overrightarrow{MV}_{ext}$ denotes the motion vector of the block that is being tested, $D_{TH,n}$ denotes the growing threshold of region n, and $D_{i,n}$ is the distance between motion vector centroid of the region and the MV of the block.

For a particular ungrouped block, if there is more than one bordering region for which (6) is satisfied, we pick the minimum $D_{i,n}$ and group the block into that region accordingly. After the block is assigned to one of its neighbouring regions, the motion vector centroid of that region gets updated as mean MV of the cluster in (5).

This step will be executed repetitively until no more blocks can be grouped into the region, and our estimate of the location of a moving object is formed.

Step-4 Update all parameters of the motion consistency model for all existing coherent motion regions, such as the motion vector centroid, the region growing threshold, and so on.

Step-5 Repeat Step 2) to Step 4) until no further seed can be found. Finally, all remaining ungrouped blocks are grouped into its neighboring region to weed out potential false motion vectors.

Step-6 Calculate the distance between centroid MVs of adjacent regions, and merge two adjacent regions if the distance between their centroid MVs is less than the threshold of minimum moving object distance, which is determined in the step 1.

Step-7 Perform region MV smoothing via vector median filtering [20] to reduce MV noise inside each region.

Fig. 7 shows the results of block-based motion segmentation with sequences Flower Garden, Tennis, and Football, where the original video frame is also displayed to give a reference on what the exact moving objects look like. Frame #2 from Flower Garden, Frame #20 from Tennis, and Frame #35 from Football are used for this example. To display the result of motion segmentation, moving regions are filled with different luminance values to distinguish them from each other.

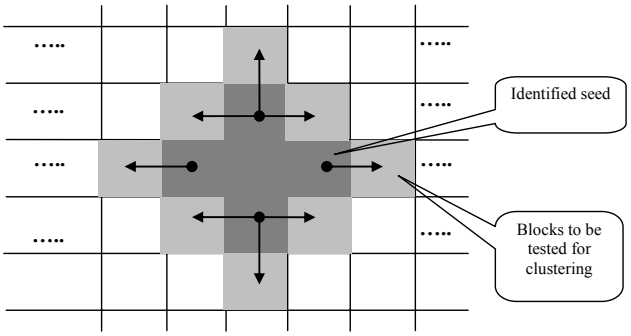


Figure 6: The region growing directions.

From these results, we can see that the proposed motion segmentation algorithm nicely segments the moving objects at a block precision level. However, it is not hard to observe that the edges of each moving object cannot exactly match its original boundary. The jagged object boundaries are caused by the block-based motion estimation at the encoder, therefore, it is necessary to come up with a method to enhance segmentation performance and subsequent MV prediction near the edges. Our solution is the variable-block-size motion prediction, which is described in next section.

2) Object-based Motion Prediction

In each coded frame, the MV associated with a block points to the most similar block in the reference frame, and can be interpreted as the motion trajectory for that block. We assume the object will keep moving in a similar direction, thus a crucial step towards synthesizing a future frame is to predict the motion for each block between the last decoded frame and the future frame. Since motion segmentation unit segments motion objects from the background area and identifies boundary blocks, we cope with them in two different ways as following.

Objects and Background: Since both the objects and the background area are a collection of unified blocks (8×8 for the case of MPEG-4), we can predict forward motion vector for either each block or entire object and background area, and median filtering is a good candidate if we go with individual block based motion prediction. In our experiment, we implement both methods, and single motion vector representation yields better results in both PSNR and visual quality.

Boundary blocks: A variable-block-size motion prediction method is employed to deal with all blocks located on the object boundary [15][21].

The blocks (8×8) are classified prior to variable-block-size motion prediction. Within each region, each block is classified as either an *internal block* or a *boundary block* in terms of its position and the energy of the prediction residual in (7), where $p_{i,j}$ denotes the prediction residual of one pixel.

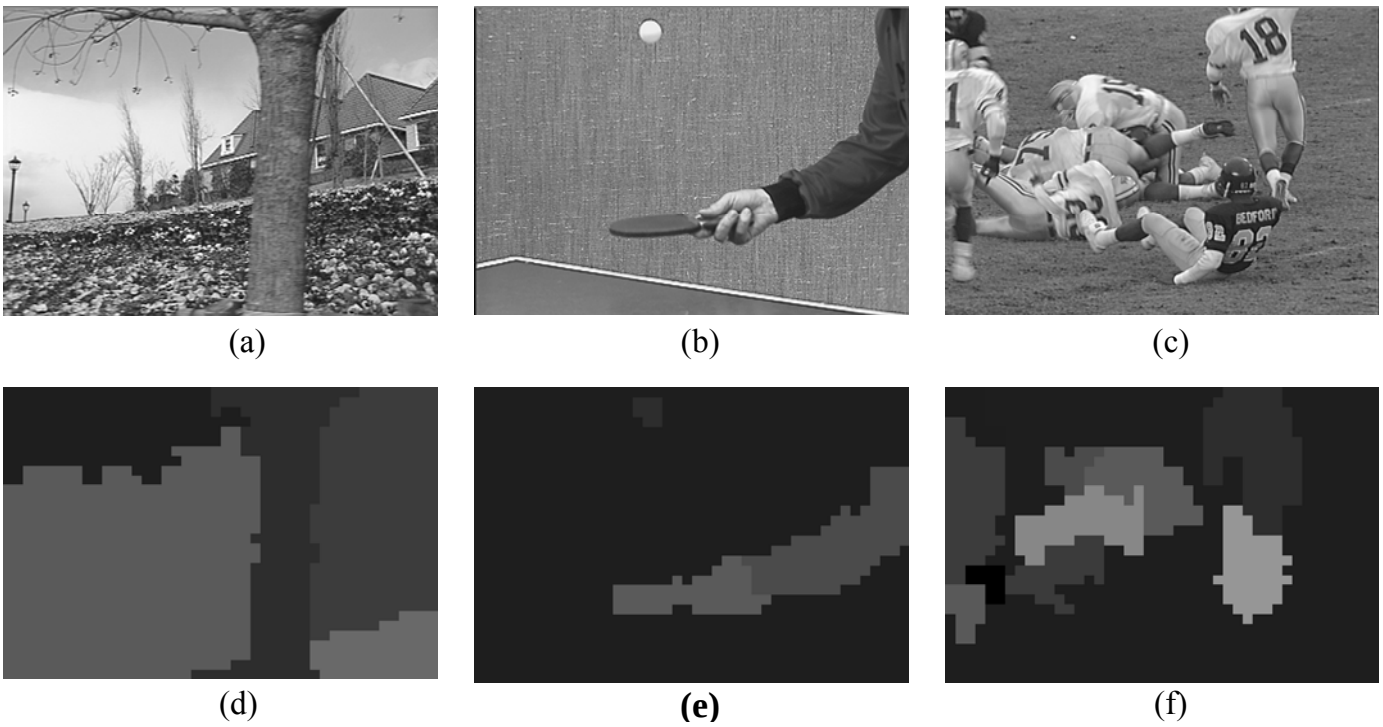


Figure 7: (a), (b), and (c) the original image from video sequences ***Garden***, ***Tennis***, and ***Football***, left to right. (d), (e), and (f) the corresponding result of block-based motion segmentation.

$$E_{res} = \sum_{i=0}^{N}\sum_{j=0}^{M}\left(p_{i,j}\right)^2, \quad N = 7, \quad M = 7; \tag{7}$$

A block is classified as a boundary block if it is located on the boundary between two motion regions and E_{res} is higher than a pre-defined threshold whose value is set to 256 based on our experiments (which indicates an average prediction error energy of 4 per pixel). Otherwise, this block is classified as an internal block.

For the block classified as a boundary block, assigning a single MV may be inappropriate since a more complex motion structure is likely involved [1]. For the case of MPEG-4, we split the 8×8 block into smaller blocks, like 4×4 sub-blocks, and predict MV for each of them to mitigate the risk of predicting a wrong single MV. Fig. 8 shows two scenarios where blocks near the boundaries of moving regions are split into 4×4 sub-blocks, and an MV is assigned to each sub-block based on which region it is closest to.

The way to assign MVs for sub-blocks is as the follows:

a) If a 4×4 sub-block is surrounded by blocks from the same region, then the centroid MV of that region is assigned to the 4×4 sub-block.
b) If a 4×4 sub-block is surrounded by multiple regions, then the MV distance is first calculated between these regions and the parent 8×8 block (before splitting). The centroid MV of the region with minimum distance to the parent MV is assigned to the 4×4 sub-block.

Let $\overrightarrow{MV}_{orig}$ be the MV of the 8×8 parent block near the boundary, and let S be a set of N motion vectors, $S = \{\overrightarrow{MV}_{cen,1}, \overrightarrow{MV}_{cen,2}, \ldots, \overrightarrow{MV}_{cen,N}\}$, whose elements are the centroid MVs of the surrounding regions. Equation (8) is used to pick the most likely MV, $\overrightarrow{MV}_{4x4,k}$, for the k-th 4×4 sub-block

$$\overrightarrow{MV}_{4x4,k} = \arg\min_{i}\left\|\overrightarrow{MV}_{cen,i} - \overrightarrow{MV}_{ori}\right\| \tag{8}$$

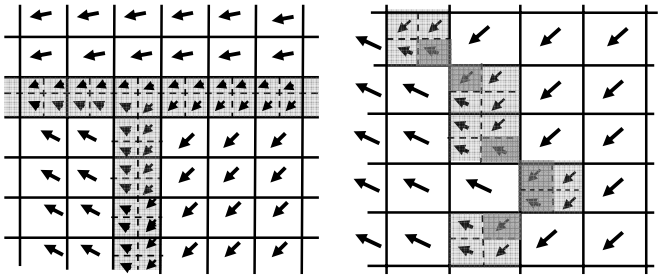
3) Preliminary Frame Synthesis

Once the segmentation is complete and MVs are assigned to each block, we move all objects and the background of the last received frame onto a future frame along the estimated MVs. In this way, we synthesize a preliminary version of the future frame. At this point, some areas of the synthesized frame may have multiple blocks landing on them – we call these areas *overlapped areas*. Other areas may remain empty, if no block lands on them – we call these *empty areas*. We need to decide which pixel values will be written into overlapped and empty areas. These decisions are made in the two post-processing steps described below.

4) Post-processing of Overlapped Areas

We distinguish two types of overlapped areas: "thin" areas are those whose width or height does not exceed 3 pixels, while "thick" areas are those whose both with and height exceed 3 pixels. Different post-processing is applied to each type of overlapped area.

For thin areas, we apply a simple averaging of all candidate pixel values. Let there be N blocks overlapping a certain area and let OV_k denote the k-th block. The pixel value at location (x, y) in the overlapped area is assigned to be the average of



A) 4×4 blocks, each of which

B) 4×4 blocks surrounded by multiple motion regions

Figure 8: The motion vector prediction for 4×4 sub-blocks.

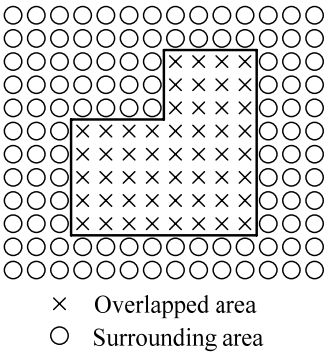


Figure 9: Boundary matching for overlapped areas.

corresponding pixel values in each of the overlapping blocks.

$$P(x,y) = \frac{1}{N}\sum_{k=1}^{N} OV_k(x,y) \qquad \textbf{(9)}$$

Once all thin overlapped areas are processed, we are left with thick overlapped whose height and width exceed 3 pixels. One such area is shown in Fig. 9.

These areas will be filled by pixel values from the block that fits the best into the surrounding area. To decide which block fits the best, we employ boundary matching by computing the mean square difference between the boundary pixels of candidate blocks, and the boundary pixels of the surrounding area.

Let $OV_k(x, y)$ be the pixel at location (x, y) in the k-th overlapping block. Let B be the set of boundary pixels of the overlapped area, and for each $(x, y) \in B$, let $n(x, y)$ be the value of the neighboring pixel across the boundary, in the surrounding area. The best matching block OV_{best} is the one whose square difference from the surrounding area along the boundary is the smallest, as in (10). Pixels from this block are used to fill the thick overlapped area.

$$OV_{best} = \arg\min_{k} \sum_{(x,y)\in B} \left\| OV_k(x,y) - n(x,y) \right\|^2 \qquad \textbf{(10)}$$

5) Post-processing of Empty Areas

In addition to overlapped areas, we also find empty areas in the synthesized frame. These are the areas where no block has landed. A similar situation arises in error concealment, where a block of size 8×8 or 16×16 may be missing due to packet loss. However, in our case, empty areas may have different

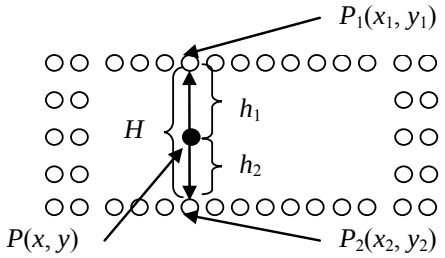


Figure 10: Filling thin empty areas.

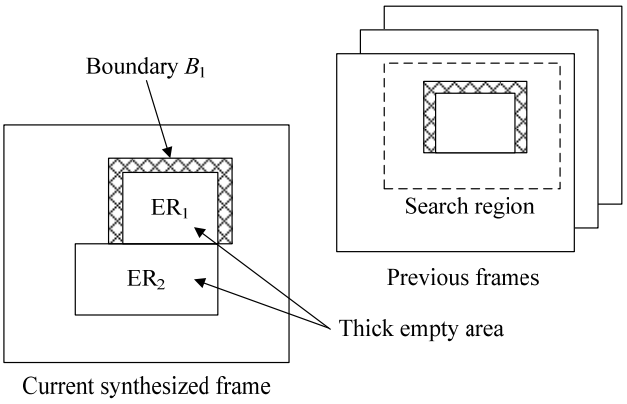


Figure 11: Filling thick empty areas.

shapes and sizes. Again, we distinguish "thin" empty areas (those whose width or height does not exceed 3 pixels), from "thick" empty areas (those whose both height and width exceed 3 pixels). Different post-processing is applied to each type of empty area.

Thin empty areas are filled using linear spatial interpolation [22]. An illustration of a thin empty area whose height is 3 pixels is shown in Fig. 10. Let $P(x, y)$ be the pixel value we wish to determine in an empty area, and let $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ be two of its nearest neighbors in the surrounding area. In the situation depicted in Fig. 10, P_1 and P_2 are above and below P, so in this case $x_1 = x_2 = x$. The pixel in the empty area is linearly interpolated as

$$P(x,y) = \left(1 - \frac{h_1}{H}\right) P_1(x_1,y_1) + \left(1 - \frac{h_2}{H}\right) P_2(x_2,y_2), \qquad \textbf{(11)}$$

where h_1 and h_2 are the distances from P to P_1 and P_2, respectively, and $h_1 + h_2 = H$.

Simple linear interpolation works reasonably well for thin empty areas, but tends to produce excessive blurring when applied to thick empty areas. Therefore, we adopt a more sophisticated method for filling thick empty areas based on boundary matching [23], [24].

An example of a thick empty area is shown in Fig. 11. First, we divide each thick empty area into rectangular regions, which we call *empty rectangles* (ERs), and label them ER_1, ER_2, …, ER_N. We fill ERs in sequence, starting with ER_1 and ending with ER_N. For each ER we extract the boundary pixels from the surrounding area and use them for boundary matching in previous frames.

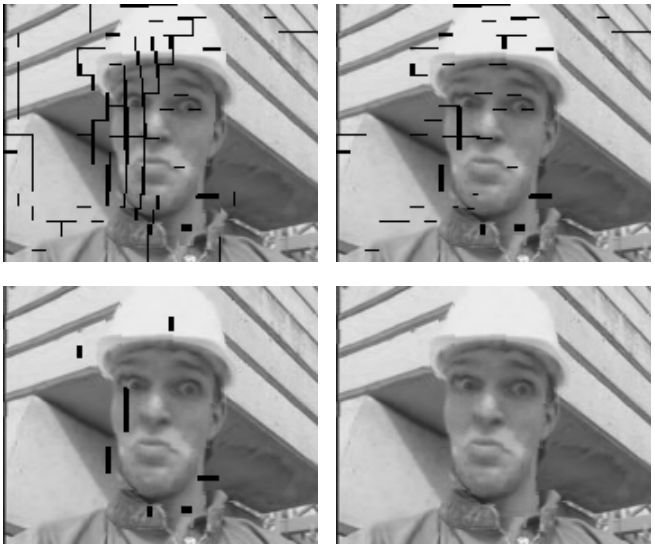


Figure 12: Illustration of empty area post-processing: intermediate frame after post-processing of overlapped areas (top left); after filling thin vertical empty areas (top right); after filling thin horizontal empty areas (bottom left); final frame after filling thick empty areas (bottom right).

Let B_n be the set of boundary pixel coordinates for ER_n. Denote the current frame as P, and previous K frames as P_1, P_2, …, P_K. We will search in each of the previous K frames over an area of size $X \times Y$ pixels for the best matching boundary. This boundary is found in frame P_k, offset by (dx, dy) from its position in the current frame, where

$$(k, dx, dy) = \arg \min_{k=1,2,\ldots,K} \min_{|dx| \le X/2, |dy| \le Y/2} \sum_{(x,y) \in B_n} \left\| P(x,y) - P_k(x+dx, y+dy) \right\|^2. \quad (12)$$

Once (11) is solved and the best matching boundary is found, we copy the corresponding rectangle from P_k to fill ER_n. At this point, ER_n is removed from the list of empty rectangles, and we continue with ER_{n+1}. The pixels of ER_n may now become boundary pixels for the remaining empty rectangles. The procedure is illustrated in Fig. 11.

An example of empty area filling is shown in Fig. 12. The figure shows a frame passing through the empty area post-processing block of Fig. 2. The top left image shows the frame produced by temporal prediction and overlapped area processing. Thin vertical empty areas are filled first (top right), followed by thin horizontal empty areas (bottom left). The final predicted frame, obtained after filling thick empty areas, is shown in the bottom right of Fig. 12.

6) Frame Prediction in H.264/AVC

Although the proposed object-based frame prediction algorithm is implemented in the XviD MPEG-4 video codec [13] for performance evaluation, it can be easily adapted to H.264/AVC [14] with similar module interface. We may expect even better prediction performance since H.264/AVC uses R-D optimal motion estimation with block size lower to 4×4 [25], [26], thus motion segmentation may be able to extract moving objects with more accurate boundaries. Variable-block-size motion prediction is still applicable to the boundary blocks by splitting them further to smaller blocks, which is very close to pixel based segmentation resolution, while the complexity will be a big consideration to the trade-off for visual quality

TABLE II: TEST SEQUENCES

Sequence	Resolution
Carphone	QCIF
Flower Garden	CIF
Foreman	QCIF
Singer	CIF
Mother & Daughter	QCIF
Miss America	QCIF

IV. RESULTS AND DISCUSSION

We have implemented the object-based frame prediction algorithm in the Xvid MPEG-4 decoder [13], and tested its performance on a number of sequences with varying motion content. We used six sequences in our experiments, each at three different frame rates: 30, 15, 7.5 frames per second (fps). These sequences are listed in Table-2. Up to 200 frames of each sequence were encoded using the IPPP… GOP structure. QCIF sequences were encoded at 128 kbps, and CIF/SIF sequences at 512 kbps.

On the decoder side, we tested prediction of up to 3 frames ahead. Depending on the frame rate (30 fps, 15 fps, or 7.5 fps) of the sequence, the perceived delay reduction is up to 100 ms, 200 ms and 400 ms when predicting three frames ahead of time. Using different combinations of frame prediction building blocks from the previous section, we constructed three prediction methods with different complexities, and compared their performance.

Method-1 is the simplest motion prediction model for frame prediction. The ZERO MV is assigned to all blocks so that there is no temporal movement at all regardless of the number of frames that need to be synthesized. In other words, the latest reconstructed frame is taken directly as the predicted frame for playout, thus it has no achievement on delay reduction itself, but it works as the basis against other two methods in terms of the prediction performance.

Method-2 applies similar block-based whole-frame concealment technique (**Algorithm-2** from section-II) [7][8] but modify it to suit multiple frames concealment. Moreover, we use vector median filter to recover MV for 8×8 blocks in the future frame, since the references of **Algorithm-2** do not detail the calculation of threshold values for MVs prediction.

Method-3 is our proposed object-based frame prediction method.

We leave out pixel-based whole frame concealment technique (**Algorithm-1** from section-II) [5][6] because it is not practical in real-time applications. The complexity becomes unacceptable especially when high-delay reduction is required (more than 100 ms at 30 fps), where 3 frames need to be predicted ahead.

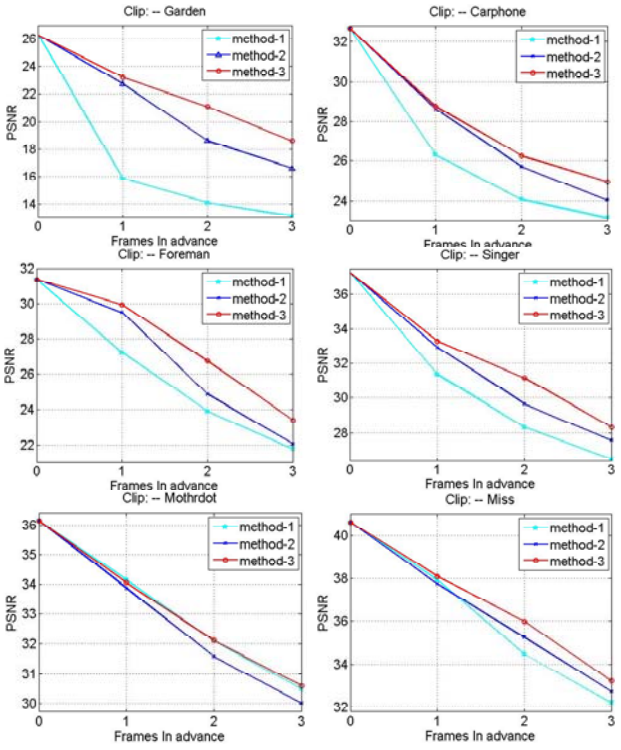


Figure 13: Prediction performance in PSNR (dB).

1) *The Performance vs. Frame Prediction Depth*

Fig. 13 shows how video quality measured by PSNR in dB depends on how far ahead we predict. The results show the frame prediction performance for different motion level sequences. All these simulations were done with 30 fps sequences.

From the PSNR plots, we can observe the followings. First, PSNR is decreasing as the prediction goes further. The quality decay varies according to the motion activity level and texture pattern in the sequence. *Garden* has the most complicated texture pattern and its PSNR (**method-3**) drops up to 3 dB when predicting one frame ahead of time, while the PSNR of **method-1** drops 10 dB after predicting the first frame. Furthermore, on sequences with relatively high motion levels, like *Garden*, *Singer*, *Carphone*, and *Foreman*, the **method-3**, proposed object-based frame prediction, outperforms both **method-1** and **method-2**. For sequences with relatively low motion, like *Miss America* and *Mother & Daughter*, the sophisticated frame prediction method does not provide as much improvement as for high-motion sequences. On these sequences, even simple methods do reasonably well. The table-3 summarizes the PSNR gain over (**method-2**/**method-3 vs. method-1)** with the prediction depth at 1, 2 and 3 frames ahead.

Fig. 14 gives a visual quality comparison between **method-2** and **method-3** for the delay reduction scenarios at 33ms, 66ms, and 99ms (Sequence *Foreman*, 30 fps). Three predicted frames, #98, #99 and #100 are predicted by **method-2** and **method-3** based on received frames up to Frame #97 respectively, and original three frames are shown on the top row for comparison.

TABLE III: METHOD-2 AND METHOD-3 PSNR GAIN OVER METHOD-1

Sequence	Predicted Frames	PSNR Gain Over Method-1 (dB)	
		Method-2	Method-3
Garden	1	6.88	7.35
	2	4.52	7.02
	3	3.49	5.5
Carphone	1	2.33	2.47
	2	1.64	2.18
	3	0.91	1.8
Foreman	1	2.28	2.73
	2	0.94	2.81
	3	0.28	1.63
Singer	1	1.56	1.94
	2	1.33	2.81
	3	1.1	1.84
Miss America	1	-0.19	0.18
	2	0.78	1.52
	3	0.53	1.01
Mother & Daughter	1	-0.27	-0.09
	2	-0.52	0.02
	3	-0.48	0.1

We can observe from Fig. 15 that the predicted frame quality deteriorates as the prediction depth increases for both methods. As expected, the further ahead we predict, the lower the quality of the predicted frames. However, the **method-3** is visibly better than the **method-2** since the block artefacts are mitigated significantly, especially for the background area, where all blocks have consistent motions.

2) *The Frame Prediction Performance vs. Frame Rate*

The frame rate is another important factor which links prediction depth and delay reduction. Figure 15 shows how the frame prediction performs at different frame rates. During the experiments, each test sequence is down-sampled from its native frame rate of 30 fps down to 15 fps and 7.5 fps. Prediction depth goes up to 3 frames ahead for each frame rate. Figure 15 illustrates the relationship between the frame prediction performance and frame rate for high motion sequences, medium motion and low motion sequences, top to bottom.

The simulation results are consistent with our expectations. Prediction is better at high frame rates. We can also observe that the quality (PSNR) of predicted frames is approximately determined by perceived delay reduction. For example, predicting one frame ahead with 15 fps *Foreman* gives us similar quality to predicting two frames ahead with 30 fps *Foreman*. In other words, the cost of quality loss to reduce perceived end-to-end delay for high frame rate source is similar to the quality loss incurred with its low frame rate counterpart.

This simulation shows that the frame prediction performance is mainly related to the amount of delay we want to reduce. However, video source with high frame rate offers us more flexibility in frame prediction than the low frame rate

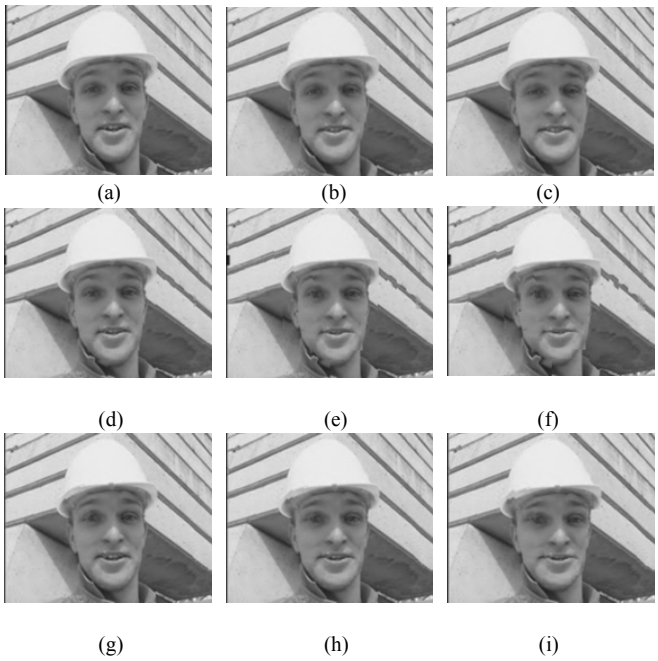


Figure 14: Visual comparison for frame prediction assessment. [Top] is the original frame #99, #100, #101 without prediction, [Middle]: frames

source. For example, if the user can accept the video quality at 150 ms delay reduction, then with 30 fps video communication, we will have options for delay reduction at 33ms, 67ms, 100ms, and 133ms, while for 15 fps video source we only have options 67ms and 133 ms, and only one option (133 ms) for 7.5 fps source.

3) *The Frame Prediction Performance vs. Motion Field Density*

Motion vector accuracy and density is another important factor affecting the overall frame prediction performance. What we have discussed so far only concerns the decoder; the encoder operates independently without any knowledge that prediction will be carried out at the decoder. In this section, we carry out experiments to investigate the impact of motion estimation at the encoder side, and how different coding techniques affect the prediction system. In particular, we investigate the effects of motion field density on prediction accuracy.

MPEG-4 supports motion vectors based on 16×16 or 8×8 blocks in terms of the energy of the prediction residual. One way to increase the density of motion vector field is to force four 8×8 blocks for inter-coding for the entire sequence, but there is a bit rate sacrifice due to the cost of encoding extra motion vectors.

The purpose of fixing the 8×8 coding mode is to increase the density of the motion vector field, which may benefit motion segmentation as well as subsequent motion prediction. However, in addition to increasing the bit rate, there is another risk associated with this. Smaller blocks may lead to more noise in the motion vectors. To identify how much improvement we can get in frame prediction in terms of motion field density and how much penalty we might suffer from video quality in terms of bit rate, we carried out the

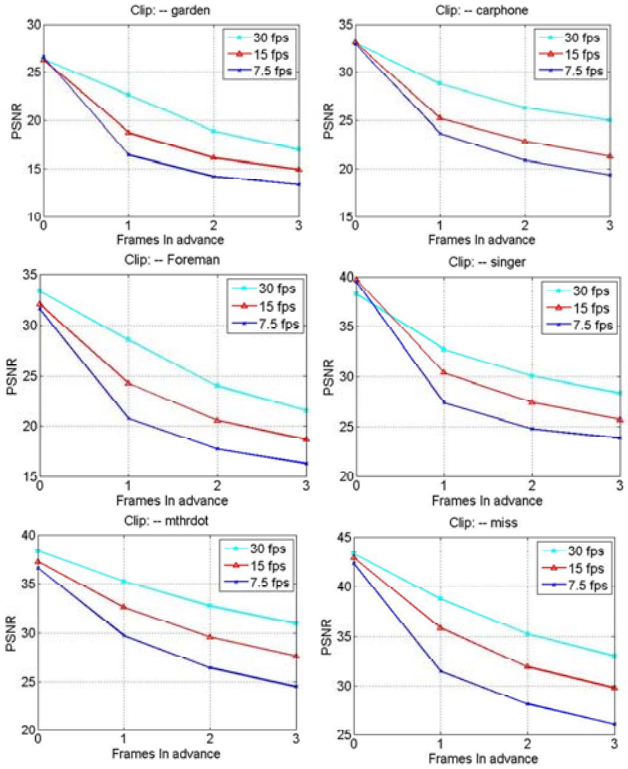


Figure 15: Prediction performance VS. Frame Rate.

TABLE IV: BIT RATES NEEDED TO MAINTAIN THE SAME VIDEO QUALITY

Sequence	PSNR (dB)	Bit-Rate (Adaptive 16×16 mode)	Bit-Rate (fixed 8×8 mode)
Foreman	31.5	70 kbps	144 kbps
Carphone	33.7	65 kbps	144 kbps
Mother & Daughter	32.4	20 kbps	149 kbps
Miss America	39.6	55 kbps	134 kbps

experiments on four sequences with different motion levels: *Foreman*, *Carphone*, *Miss America*, and *Mother & Daughter*. Two compressed video files are used for frame prediction testing at the decoder separately.

- Fixed encoded video quality (PSNR) with different bit rates.
- Different encoded video qualities (PSNR) at the same bit rate.

Table 4 shows the bit rate difference when we use the two different coding modes to achieve the same quality (PSNR), and we can observe that the adaptive 16×16 mode is more efficient that the fixed 8×8 mode. Efficiency difference between these two modes is highest on low-motion sequences, where lack of motion can be easily exploited by the large block size. The difference in efficiency reduces as the motion intensity and complexity increase.

Fig. 16 illustrates the frame prediction performance in terms of different motion field densities for four different video sequences. The plots on the right side correspond to the case where the encoded PSNR is the same for the two modes, and

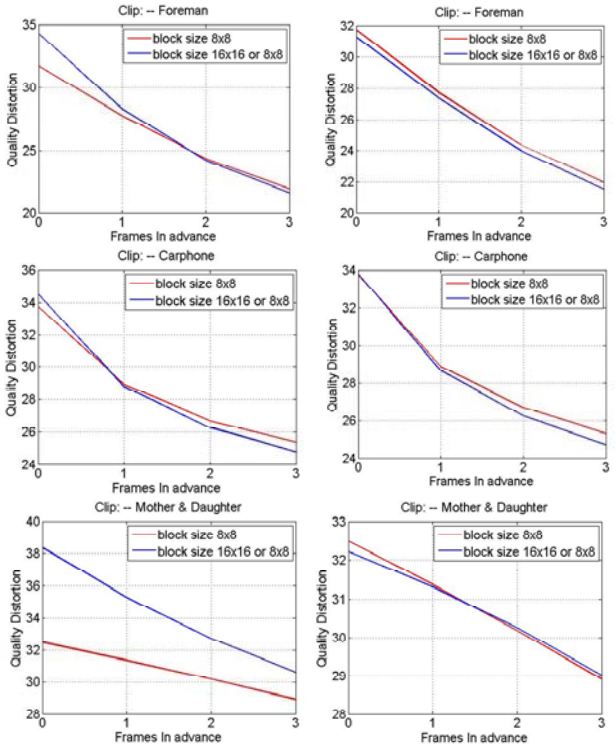


Figure 16: Prediction performance vs. Motion density.

TABLE V: BIT RATES NEEDED TO MAINTAIN THE SAME VIDEO QUALITY

Video Clip	Format	Time Consumption - ms			
		STD. Dec	1 ahead	2 ahead	3 ahead
Foreman	QCIF	0.26	1.93	4.74	6.72
Carphone	QCIF	0.16	1.92	3.69	6.41
Singer	CIF	1.38	6.6	13.9	21.5
Mother & Daughter	QCIF	0.42	1.35	3.28	4.69
Flower Garden	CIF	2.87	8.84	17.95	27.3
Miss America	QCIF	0.61	1.87	3.33	5.42

TABLE VI: THE MIPS DISSIPATION.

Processing Unit	MIPS percentage
Motion Segmentation	40%
MV. Pred	15%
Frame Synthesis	29%
Post Pro - Overlapped	5%
Post Pro - Empty	4%
Others	7%

the plots on the left side correspond to the case where the bit rates of the two modes are the same.

We can observe that increased motion vector density does not give too much improvement on the performance of frame prediction, especially when encoded bit streams have the same PSNR. The main reason is that increased motion density brings higher risk of false motion vectors in homogeneous areas, and motion accuracy turns out to be more important than motion field density in frame prediction.

By studying the impact of motion field density on frame prediction, we have come to the conclusion that increasing the motion field density by forcing small block size in motion estimation is not beneficial for predictive decoding. Using the default adaptive 16×16 block-based motion estimation produces less dense motion field, but has the benefit of improved coding efficiency and higher motion vector accuracy, which is crucial for prediction.

4) The Complexity Analysis

The complexity has been a big practical consideration in frame prediction system. We analyze the computation complexity by conducting simulations on a desktop PC with Intel Pentium CPU 3.0 GHz and 1.99 GB of RAM for the proposed frame prediction algorithm (**method-3)**, and table 5 gives us a preliminary knowledge of algorithm complexity with an un-optimized version of the frame prediction module, where the overall time consumption is listed for standard decoder and multiple frame prediction. Table 6 shows the computation dissipation in current implementation.

To achieve real-time video decoder with frame rate at 30 fps, current frame prediction implementation can support the video resolution up to CIF, while predicting three frames ahead of time, and reducing delay up to 100 ms. For lower resolution such as QCIF, the overall time consumption, including frame reconstruction and three-frame prediction, is less than 10 ms, thus it meets the requirement of running in real time. However, if a real-time encoder needs to run on the same system (as in typical PC-based videoconferencing), the current frame prediction module would need further optimization to reserve enough MIPS, especially for higher resolutions like CIF and longer delay reduction.

V. DISCUSSION AND CONCLUSION

In this work, we addressed the issue of delay in video communications, and proposed predictive decoding to reduce the delay. The main idea behind our delay reduction method is to incorporate a frame prediction module into a standard video decoder, predict the upcoming video frames from the available video data, and display them before they arrive at the receiver.

By using this "predictive decoding," it is possible to reduce the perceived end-to-end delay at the expense of displayed video quality. The key steps in this decoding scheme include:

a) Motion segmentation-based motion vector prediction, which considers both prediction residuals from the last received frame and motion region homogeneity to improve motion prediction,
b) Temporal prediction is used to synthesize future frames using the predicted motion vectors,

c) A combination of bi-linear interpolation and boundary matching is used as a post-processing step to deal with overlapped and empty areas.

The benefits we may get from frame prediction depend on the complexity of the video. For scenes involving complex motion, the number of frames which can be predicted with acceptable quality is very limited. However, for scenes with less complex and relatively smooth motion, as we might expect in a video conference, predictive decoding shows promising performance.

REFERENCES

[1] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video Processing and Communications*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[2] J. A. Zebarth, "Let Me Be Me [Video Telephony and Teleconferencing Tests]," *Proc. IEEE Globecom'93*, Vol. 1, pp. 389-393, Nov. 1993.

[3] S. Desmet, B. Deknuydt, L. Van Eycken, and A. Oosterlinck, "Classified Motion Estimation for Video Coding," *Proceedings of SPIE*, vol. 2182, Image and Video Processing II, pp. 111 – 119, March 1994.

[4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video coding Standard," *IEEE Trans. Circuits Syst. Video Technol.*, , vol. 13, no. 7, pp. 560-576, July 2003.

[5] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, "Concealment of Whole-Frame Losses for Wireless Low Bit-Rate Video Based on Multiframe Optical Flow Estimation," IEEE Trans. Multimedia, Vol. 7, No. 2, pp. 316-329, April 2005.

[6] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, "An Error Concealment Algorithm for Streaming Video," in *Proc. ICIP-IEEE int. Conf. Image Processing*, Vol. 3, pp. III-649-52, Sept 2003.

[7] P. Baccichet, D. Bagni, and A. Chimienti, "Frame Concealment for H.264/AVC Decoders," *IEEE Trans. On Consumer Electronics*, Vol. 51, Issue 1, pp. 227-233, Feb 2005.

[8] P. Baccichet and A. Chimienti, "A Low Complexity Concealment Algorithm for the Whole-Frame Loss in H.264/AVC," *in Porc. IEEE MMSP'04*, pp. 279-282, Oct 2004.

[9] E. Quacchio, E. Magli, G. Olmo, P. Baccichet, and A. Chimienti, "Enhancing Whole-Frame Error concealment with an Intra Motion Vector Estimator in H.264/AVC," *in Proc. ICASSP-IEEE int. Conf*, Vol. 2, pp. 329-332, March 2005.

[10] S. C. Sui, F. Liu and P. Xue, "Whole-Frame Error Concealment with Improved Backward Motion Estimation for H.264 decoders," *IEEE fourth International Conference on Image and Graphics*. pp. 279-284, August 2007.

[11] Y. M. Chen and I. V. Bajić, "Predictive Decoding for Delay Reduction in Video Communications," *Proc. IEEE Globecomm'07*, pp. 2053-2057, Washington, DC, Nov. 2007.

[12] *Information Technology - Coding of Audio-Visual Objects, Part 2: Visual*, ISO/IEC 14496-2, 1999.

[13] XviD MPEG-4 codec: http://www.xvid.org/

[14] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Recommendation H.264, Mar. 2005.

[15] B. D. Choi, J. W. Han, C. S. Kim, S. J. Ko, "Motion-Compensated Frame Interpolation Using Bilateral Motion Estimation and Adaptive Overlapped Block Motion Compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, Apr. 2007.

[16] J. Y. A. Wang and E. H. Adelson, "Representing Moving Images with layers," *IEEE Trans. Image Proc.*, vol. 3, no. 5, p.625-638, Sep. 1994.

[17] G. D. Borshukov, G. Bozdagi, Y. Altunbasak, and A. M. Tekalp, "Motion Segmentation by Multistage Affine Classification," *IEEE Trans. Image Processing*, vol. 6, no. 11, pp. 1591-1594, Nov. 1997

[18] T. Kanungo, D. M. Mount, et al., "En Efficient k-Means Clustering Algorithm: Analysis and Implementation," *IEEE Trans. On Pattern Analysis and Machine Intelligence* vol. 24, no. 7, pp. 881-892, 2002.

[19] H. Y. Chung, Y. L. Chin, K. Wong, K. P. Chow, T. Luo, and S. K. Fung, "Efficient Block Based Motion Segmentation Method using Motion Vector Consistency," *Proc. IAPR Conference on Machine Vision Application*, pp. 550 – 553, May 2005.

[20] J. Astola, P. Haavisto, and Y. Neuvo, "Vector median filters," *Proc. IEEE*, vol. 78, no. 4, pp. 678-689, Apr. 1990.

[21] M. Silveira, M. Piedade, "Variable Block Sized Motion Segmentation For Video Coding," *IEEE ISCAS*, June 9-12, 1997.

[22] A. Katsaggelos and N. Galatsanos, *Signal Recovery Technique for Image and Video Compression and Transmission*, Kluwer Academic Publishers, 1998.

[23] M. J. Chen, L. G. Chen, and R. M. Weng, "Error concealment of lost motion vectors with overlapped motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 3, pp. 560-563, Jun. 1997.

[24] Y. Chen, O. C. Au, C.-W. Ho, and J. Zhou, "Spatio-temporal boundary matching algorithm for temporal error concealment," *Proc. IEEE ISCAS*, pp. 686-689, May 2006.

[25] J. Stottrup-Andersen, S. Forchhammer, and S. M. Aghito, "Rate-distortion-complexity optimization of fast motion estimation in H.264/MPEG-4 AVC," *Proc. IEEE ICIP'04*, Vol. 1, pp. 111-114, Oct. 2004.

[26] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-Constrained Coder Control and Comparison of Video Coding Standards," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 13, No. 17, pp. 688-703 July 2003.

Yue-Meng Chen received the B.S. in electrical engineering from Zhejiang University, Zhejiang, China, in 1998 and the M.A.Sc in engineering science from Simon Fraser University, Burnaby, BC, Canada in 2007. He is currently pursuing the Ph.D. degree in engineering science at Simon Fraser University, Burnaby, BC, Canada.

From 1998 to 2000, he was a DSP engineer with Hualong information and technology development centre, Shanghai, China. From 2000 to 2001, he was with Future Electronics Inc, Shanghai, China. From 2001 to 2007, he was a senior DSP staff engineer with Centrality Communication Inc, Redwood city, CA, USA. Since 2007, he is a senior GPS staff engineer with SiRF Technology Inc, San Jose, CA, USA. His research interests include multimedia communications, GPS technologies in signal acquisition and tracking.

A Framework for Interdisciplinary Interactive Performance through Perception of Intention

Greg Corness
School of Interactive Arts and Technologies
Simon Fraser University
1345 King Edward Highway
Surrey B.C. Canada
gcorness@sfu.ca

ABSTRACT
Interactive computer systems have been incorporated in the performing arts for the past two decades. During this time, several approaches to the performer-technology relationship have been developed. However, resent increases in agency and autonomy designed into performance systems challenges traditional view on Human-computer relationships. This study presents the foundation for an interaction framework, by drawing on theories from cognitive science, performing arts and philosophy. These theories argue that the embodied perception of the intention of others provides the basis for human interaction. Therefore, the human-computer interaction model presented focuses on the performer ability to perceive the intentions of the system

Keywords
Interactive performance, Perception, embodiment, HCI

1. INTRODUCTION
Interactive technologies have been use in performance for the past two decades; but in resent years, explorations into the use of these devices have expanded. Music, dance, and theatre have individually developed models for incorporating and interacting with media and digital systems onstage. The increase in availability and usability of interactive technologies has allowed practitioners to become more comfortable and adept at incorporating these technologies into productions, there by encouraging an exploration into the use of these devices and the complexity of the performer-system interaction. Interactive performance is pushing the explorations in human-computer interaction from pure technical questions to philosophical questions. Theorist and director David Saltz remarked in his essay *Live Media: Interactive Technology and Theatre* "The holy grail for me as a director is to produce a dramatic relationship between performer and media, to grant media real agency, casting them in a role on par with the live performers" [15]. Computer systems are being given increasing autonomy, our relationships with these systems are increasing in complexity, and our artistic questions are challenging the divide between performer and system. This essay focuses on the relationship and interaction between performers and digital systems in order to develop a framework for interdisciplinary interactive performance.

The proposed framework expands on traditional approaches to interactive performance by re-conceptualizing the media object, shifting it from a virtual object to a perceived body within the physical space. The re-conceptualization of the media object is guided by a theory of embodied knowledge, which illustrates how *perception* acts as the key to the performer's knowledge about the environment. The embodiment of the media object through the human process of perception constructs the media object as a lived experience equal to other physical experiences. This perceptual approach to interaction is supported by resent research in cognitive science and theories informed by performance practice. Because this experience is rooted in the performer, I refer to the framework that I am developing out of these theories the *Performer model*. The Performer model for interaction places the quality of the Human-Computer relationship on the ability of the performer to perceive the intentions of the system. The intent of this essay is to present how the human ability to perception the intention of others acts as the foundation for this framework.

1.1 Taxonomy of relationship
With in the disciplines of dance, music and theatre, several approaches for incorporating interactive systems in stage performance have been developed. Each approach affords a different function to the technology and places it in a different relationship with the performer. Saltz has presented a taxonomy of these performer-technology relationships [15]. His list of relationships consists of:

1) Virtual Scenery: - media that provides a static or animated backdrop to the performer's actions.
2) Interactive Costumes: - media embedded in the costuming to provide an expressive quality that is responsive to other performance elements.
3) Alternate Perspective: - media used to present an alternative physical perspective of actions seen on stage.
4) Subjective Perspective: - media used to present the internal thoughts of a character on stage.
5) Illustration: - media used to illustrate a performer's words.
6) Commentary: - media providing supplementary commentary on the stage action in the moment of the action.

7) Diegetic Media: - media within the world of the performance, acknowledged by the performers as media.
8) Affective Media: - media not in the world of the performers, provided to guide the emotions of the audience.
9) Synesthesia: - media reflecting the state of the performance in a second sensory modality.
10) Instrumental Media: - media acting as an extension of the performer with an arbitrary, though consistent, relation between the performer's actions and the system's response.
11) Virtual Puppetry: - media used to present a performer's double or alter ego. The performer gives their voice to the puppet.
12) Dramatic Media: - media objects functioning as characters within the drama of the performance

This extensive taxonomy, largely informed through the perspective of theatrical function, provides an overview of the relationships between the performer and the media. The organization places media objects supplying context, media supplying uninvolved commentary, media translating performed gestures, and media incorporated as agents within the performance on a continuum of Performer-Technology relationship. The discussion around each category focus on the role the media serves, and the affordences brought to the role by the media.

However, as much as Saltz has illustrated the range of system-performer relationships on stage, his taxonomy is hampered by the difference between *relationship* and *interaction*. In explaining the concept of *Dramatic Media*, a media functioning as a character, Saltz uses the example of a live performer playing alongside a video of a performer, describing both as characters in the scene. What Saltz is presenting is a dramaturgical relationship, not an interaction. Both characters have agency in the performance only from the perspective of the Director and author. Their status as equal characters is constructed by focusing on their abstract agency in the composed structure. The description of their relationship fails to address the presence or absence of any interaction enacted in the moment.

1.2 Taxonomy of interaction

An alternative approach to exploring the performer-technology relationships is to consider the mode of interaction afforded by the system. Robert Rowe proposed the distinction between an instrument paradigm and a player paradigm as one axis along which we could place different interactive systems [9]. In this taxonomy, systems based on the Instrument paradigm are often more concerned with timbre generation, while the systems modeled on a player paradigm requires the use of some meta-compositional generation method to produce musical output [14].

By focusing on the process used to generate the system's response, Rowe draws a distinction between two fundamental modes of interaction: direct and indirect. The direct control associated with the instrument paradigm implies an interaction based on one-to-one mapping between the performer's action and the system's response, often focusing on control of musical parameters such as on/off, pitch, and volume. In contrast, the indirect control associated with the player paradigm implies an interaction based on sensors and general Performative gestures, analyzed and mapped to one or more *compositional* parameters [19]. These two paradigms afford the performer and the system very different levels of control, servitude and freedom. The performer-system relationship becomes defined by the analysis and mapping strategies being employed, which are consequently understood as key structural qualities of the system and the performance [13] [19]. A dynamic quality to the interaction may even be constructed by change the mapping and analysis strategies in a set manor over time [2], adding a further dimension of complexity to the relationship. However, the link between mapping strategies and control based paradigms limits the usefulness of the instrument-player taxonomy to fully contend with issues of agency. Rowe's use of the term "player" and Saltz's use of the term "character" implies that agency is plays a role in the discussion; however, neither Saltz nor Rowe address, in these taxonomies, how the interaction or relationship with such a system might be negotiated.

1.3 Models for improvised interaction

The agency of the media performer is a common issue in research with computer systems for musical improvisation. Interaction with these systems is often facilitated by elaborate algorithms for analyzing the sonic gestures of the performer [9] [16], which have been modeled on the process assumed to be undertaken by the performer. This model has some validity and has produced some very accomplished systems, and seems to agree with a call-and-response information transfer paradigm of interaction. However, the analysis model for interaction has focused research on the computationally difficult task of syntactical understanding and response. The assumption being that the communication requires the transfer of information in the performed medium, ignoring the general social connection between performers. Generally, performers remain in what might be termed *social contact* with each other, by placing themselves with in line of sight and/or range of other communication modes distinct from the sounds they are producing. This preference, shown by most performers, suggests that these other forms of communication affect the performer's ability to interact the others.

Interactive performance projects, largely in the context of interdisciplinary performance, have already been exploring a variety of communication cues such as visual movement, acoustic variation [9] [18] and multisensory multimedia [3] to expand the sensory relationship between the performer and the system. Explorations in Multimedia and interdisciplinary interaction have indicated that the performer communication cannot rely on the syntax of a particular performance domain, since the communication must cross between domains such as music and dance. Often such interactions become dependent on general expressive gestures that are not domain specific and may be associated with pure human body expression.

2. EMBODIMENT OF MEDIA OBJECTS

If we are to put the physical body in an interactive relationship with a media object, an understanding of the perceived body is required. Writers such as Heidegger, Merleau-Ponty, and Varela have put forward the idea that all knowledge of the self and the world starts with the lived experience. This knowledge includes the *body knowledge* understood to reside in the body such as muscle memory, habit, and instinct, but has at its core the process of *perception* in which the body and mind are unified in the act of sensing, and understanding the world that we are a part of [12]. Implicit in this understanding of knowledge is the idea that there is no separation between the physical body and the cognitive self, no Mind-Body separation. The key concept is that I am part of the world. I am neither in the world (empiricism) or have the world in me (intellectualism) but rather that I am both subject and object at the same time in constant interaction. Merleau-Ponty describes this state as being both a thing among things and that which sees and touches [11]. He refers to this idea in his philosophy of embodiment as *Reversibility*.

Further to the notion of reversibility, Merleau-Ponty maintains that our knowledge of the world is constructed by the interplay between our presence in it and our perception of it.

> "It is the body and it alone, because it is a two-dimensional being, that can bring us to the things themselves, which are themselves not flat beings but beings in depth, inaccessible to a subject that would survey them from above, open to him alone that, if it be possible, would coexist with them in the same world" [11].

When we consider that our own body is both an objective body and a phenomenal body, that constructs the world by being part of it [11], then perceiving the body of the digital object is a matter of perceiving an experience with it through our body senses. The body of the object is constructed through our embodied experience with it. Consequently, *body* may be viewed as a perception of possible interaction more then a physical state.

In a media installation, *Arranging With a White Rabbit*, I explored this approach to knowing the body of a media object. The installation involved a moving light, a virtual windchime and a musical trill (a musical gesture independent of any instrument). These three objects illustrated three approaches to abstraction and space. The light defined area visibly with a constant lighting texture even though it moved throughout the space. The windchimes were invisible but defined area sonically through a consistent sound texture that remained spatially located. The trill defined a musical space, existing only as the musical gesture of a pitch change and rhythmic texture. The trill remained consistent in its gestural space, but moved throughout the virtual ensemble, allowing it no consistency of timber or physical association. During the studies that led up to the installation, I discovered that that all three of these media objects could be perceived as embodied objects in the physical space. Through the sensual perception of the system's responses to gestures and objects in the space, the experiencer's knowledge of the space was constructed to include knowing the Media objects as responsive embodied object inhabiting the physical space. One key factor to this knowledge was that the object's response, constructed by the system, needed to be identifiable with the object. Associating the response with the object was easily achieved by designing consistent responses, and yet, a strictly consistent response defined the objects perceptually as inanimate objects, not as agents acting in the space.

Still, the objects were perceived as bodies in the physical space, regardless of the media and abstraction used to define and render them in the physical space. The study shows that through experience, our senses provide input about the world, but the input is not knowledge until it is *perceived* and built up into a worldview. Perception is the reflective process involving all the data being sensed including sensation of self-knowledge. The self-knowledge gives context that makes the perceived input knowledge. It is the interaction between self-knowledge, sensation, and perception that forms our knowledge and our world.

3. IMPROVISATORY PERFORMANCE

The understanding of embodiment and embodied knowledge also has implications in understanding the communication process between performers. Communication in performance is an inter-subjective phenomenon, where understanding is *agreed* upon by the agents involved in the moment. As Lockford and Pelias explain:

> "Even when faced with the challenge to perform in an unscripted moment, performers understand that they are engaged in an ongoing communicative exchange. This exchange is a process best conceived, not as an act of information transmission or shared understanding, but as communication scholar H. L. Goodall, Jr. would have it, as an act of 'boundary negotiation'." [10]

Here "boundary negotiation" refers to the process in which the self of the performer is incremental built with in the context of the performance. This notion of building is parallel to the notion of the self being both subject and object in process of constructing knowledge. In a theatrical sense, this process is the build up of character and location as new information is reveled during the scene. Such a negotiation implies that the agent must be able to respond to new information while simultaneously presenting information to contribute to the self of other agents. Negotiation in these terms is a coordination of the interaction between agents [10]. Based on this requirement to coordinate their interactions, it becomes imperative that all agents are able to negotiate the coordination of their intention and therefore able to track the intention of the others.

The importance of the agent's ability to track the intention of others is highlighted further through considering issues of trust. Since the agents constitute themselves and each other through the negotiation of boundries [10], this inter-subjective communication requires a sense of trust. For a performer to be open to constituting their performance identity anew in negotiation with others on the stage, they

must trust the environment. Furthermore, a sense of support is established when their actions both affect and support other agents. The support builds out of the trust in the inter-subjective understanding of the moment, the trust that all agents' understanding of the moment points to the same goal. Keeping the ensemble synchronized is a trust in the inter-subjective understanding, and requires all the performer-agents are able to track the intention of the others. Consequently, all agents need to be able to project their intentions.

4. AGENCY

Trust has further implication on agency. Bogart and Landau coach students of theater improvisation to "*trust* in letting something occur onstage, rather then making it occur" [1]. Although this statement implies that the performer's agency should be initiated, it does not mean that nothing should be started, but implies an avoidance of forcing a start. We might call this *letting something occur* an additive approach to interaction where additive suggests that the agency is added to the state of the system regardless of the system's current state. The notion of adding has strong implications for the traditional response communication model. To trust in *the something that will happen* is to coordinate actions in order to add to actions that are happening. The act of coordination cannot be done in response. The improviser must move beyond the cognitive and trust in the intuitive [10]).

5. INTUITION AND INTENTION

Research in the field of neuroscience has recently suggested links between *intuition* and *intention*. Neurons found in pre-motor areas of the brain have been shown to fire, not only when producing a sound or action, but when the subject hears the sound or observes others performing the same action as well [5]) [7] [6]. The firing of these neurons allows the subject to predict the outcome of their own actions as well as the actions of others. "This implicit, automatic, and unconscious process of motor simulation enables the observer to use his/her own resources to penetrate the world of the other without the need of theorizing about it" [4]. What is crucial to the phenomenon is that the action observed must be goal oriented; in other words, it must have intention [5] [7] [6]. These findings suggest that as social being we have developed the ability to intuitively predict the actions and sounds of objects and agents around us.

Still, the research leaves some question as to the usefulness of mirror neurons in human-computer interaction. The findings to date, concerning a person's ability to perceive intention in others, suggest that the ability diminishes in correspondence to the physical similarity with the other. The implication of these findings is that a human subject perceives, in a pre-conscious manor, the intention of other humans, but less so the intention of apes, only slightly those of other animals. These tests also suggest that there is no pre-conscious perception of the intention of a machine. The prevalent reason given for these results is the degree of movement similarity between subjects [5]. Thus, it is unclear whether the system's actions/response, when modeled on the response of a performer, would affect the per-cognitive process of a human performer.

However, the presence of the pre-cognitive function implies that the cognitive system as a whole works in connection with the pre-cognitive mechanism; and that even at a cognitive level, interaction is governed by the prediction of events as much or more then the reaction to events. The claim that animals, including humans, respond to the becoming of an event is also supported by David Lee's General Motion Gap theory. His theory addresses the cognitive process undertaken by any animal coordinating its actions with its environment. In his paper *How Movement is guided*, Lee suggests that the information guiding movement must be extrapolated into the future in order to facilitate the rapid guidance of movement [8]. In his theory, Lee discusses interactions with both physical objects and metaphysical objects such as musical pitches. This theory implies that an animal's interaction with an intended goal, physical or virtual, is guided by a cognitive process involving prediction.

The separation of time and action is also evident in research done on conscious will and motion. Of special interest are the indications that *intention* happens significantly before the action is physicalized. Wegner, in his book "The Illusion of Conscious Will", presents the work of Kornhuber and Deecke (1965) as well as Libet (1983). These researchers ran studies showing that an increase in brain activity occurred in subjects up to 800ms before a voluntary action took place. In the case of Libet's experiments, brain activity was recorded over 300ms before the subject was even aware they wanted to act [17]. These findings further indicate that humans do not live in a static present moment, but rather in a moment becoming the next. Our social engagements are informed by an embodied empathy that allows minor predictions of those around us. We react not in the moment but in the moment next over half a second late.

6. PERFORMER MODEL

These theories provide and argument that humans interact with the world by perceiving the intentions of both themselves and others in the environment. This argument supports the presented theory on communication in performance, which proposes that the communication between performers be understood as a negotiation and coordination between agents. Based on these theories, I suggest that a framework for interaction must address:

1) The need to negotiate boundries and build trust with others
2) The development of an inter-subjective understanding of the moment
3) Agency and acceptance in the environment

The words *Trust*, *Understanding* and *Acceptance* are crucial here, but not because of what they mean. The association of these words with human experience indicates that the focus of the framework must be the performer's experience. And yet, embodiment theory argues that experience is based on perception, and the theories presented from cognitive science and performance practice demonstrate that perception of intention is key to our interactions with our environment and the agents in it. Through this line of reasoning, we may now suggest that perceiving intention is integral to the performer

experiencing of agency, acceptance and trust. Therefore, the foundation to the proposed framework is an understanding of the mechanism that allows the performer to perceive the system's intention. The following studies have been a step towards this understanding.

6.1 Testing the Frame

I performed a series of studies looking at the experience of the performer working in a system designed to project its intention. The system used to conduct the studies took two forms, visual and sonic. Both systems were designed using an iterative process guided by a first person methodology. In order to focus the study on the perception of the system's projected intention, gestures in both systems were generated using simple random processes. The use of randomly generated gestures allowed the system to enact its own "intention" with no control or other data of the performer, and avoided any signifiers associated with expectation, structure or syntax in the medium. The response paradigms chosen for both modes were informed by human response and perception behavior but were not intended to mimic it. Finally, the research was set up as studies into the experience of being afforded the ability to perceive the system's intention. No expectation of creation or performance was imposed.

The response gestures in the visual system were realized using an image of two concentric circles projected onto the floor of the performance space using an I-CUE dmx controllable mirror. The behavior of the system was set so that the inside circle needed to move off center for the entire image to move in the space. Stopping required the circle to return to the center. The direction and amount that it moved off center corresponded to the direction and speed at which the image was about to move. Since the movement of the inner circle was analogous to the thinking and preparing to move of a performer, the duration set for the inner circle to reach its maximum point was 300ms, in line with the research presented by Wegner. The speed at which the light was about to move was mapped to the distance moved by the inner circle off of center, modeled loosely on a bodies need to shift its center of gravity in preparation to move. The movement of the light object was constrained using a dynamic weighted random algorithm. The probability of the light moving in any direction was a function of its position in the space.

The sonic version of the study developed from the common idea that breath acts as an embodied connection within a group. The acoustics of a flute were digitally modeled, allowing the synthesis of the breath associated with playing the instrument to be controlled separately from the tone of the flute. Using the digital model as the sound source, a number of sessions were conducted, exploring the possibility of perceiving information concerning the *intention* of the system. The information was embedded in different parts of the breath sound by manipulating the parameters of the flute model to control the quality and duration of the breath sounds before and after the flute tone. The timings of the different sound qualities were functions of the generated gesture's length, density and speed.

6.2 Experiencing the system

I spent a number of sessions working in the two systems to feel the experience of being in the space *with* each system. As might be expected, anthropomorphizing the light was almost automatic. I perceived its motion as a nervous exploring intention, even though I *knew* the movements were purely random. And still, as much as I perceived the motions as exploratory, I quickly became aware that the system had no sense of my presence. Although this limitation had been part of the design, it was interesting to note how easily I perceived the design as experience. In fact, though the model for the interaction design had been "tag", due to its lack of awareness and my ability to anticipate its movements, the interaction quickly became like playing in ocean waves or taunting a blindfolded partner. The random movements had a lunging and staking quality. But without the impression of being the object of it's stalking, I noticed myself considering its projected intended motion and circling to the other side just out of "reach". The random process used for starting and stopping also produced occasional "fakes" in which the Light Actor moved its "weight" in one direction then immediately moved it back to a center position. This emergent behavior was of special interest. The perception that I could tell where it was "thinking" of moving encouraged me to get close but the impression that it could "change its mind" kept up the engagement.

During a second session, I focused on moving with the light rather then on avoiding the light. At first, I changed only my behavior; the system's behavior pattern remained the same as before. My experience of this interaction was very unsatisfying. Although I could tell where the light was going, I had very little time to coordinate my own movements. The interaction quickly became a dodging rather than a *moving with*. I therefore changed the systems behavior settings to generate movements that tended to be longer with less "fake" motions. These changes were modeled after *mirroring exercises* in which human partners try to mimic each other's motion without a sense of leading. In these exercises, fluid, often slow, predictable motions are emphasized. With the system's behavior modeled on *mirroring exercises*, I found the experience of moving with the light became possible; however, the quality of my movement remained at a "proof of concept" level. The interaction did not inspire flow or exploration in my own movement. I could *move* with the light but not *dance* with the light.

The audio-based system had a much different impact. Whereas the visual system had initially inspired an avoidance response and only after being re-modeled, produced a *moving with* response, my experience was that the breath model in the audio based system immediately inspired a *moving with* response. The breath sound in the first session was linked to the duration of the generated phrase and produced a feeling of lift into tone of the sound. The feeling of "lift" encouraged the alignment of my motion *with* the onset of the sound even though I had no knowledge of when it would happen.

Through reflecting on my response, I noticed that I perceived the breath generated by the physical model in two distinct

parts: the inhale and the stream focusing. I also noticed that I was preparing to move on the inhale and initiating the moving on the small change of breath sound just before the flute tone. This discovery inspired a series of sessions exploring the breaking of the breath sound into three parts: inhale, focused air-stream, and breath trail-off. I then mapped inhale duration and volume to phrase tempo and duration respectfully, based on the experience that breathing into a beat often signals a down beat and that more air is needed to play longer phrases. The system using this metaphor allowed me to often anticipate the random phrase and move with it in tempo, but only within a small range of tempo values. However, when inhale duration was calculated as a function of phrase length, I found that I moved with the sound without much thought. Through reflection of this observation, I concluded that model linking duration of breath to tempo was experienced as a culturally learned sign. This realization illustrated the issue of cognitive load in interpretation as a concern for any system that is based on perception.

7. CONCLUSION

Humans have developed a complex process for generating knowledge through lived experience. The philosophical theory of Embodiment describes this process as interplay between our presence in the world and our perception of it. This theory converges with theories from performance practice, which suggest that human interaction is a process of perceiving and building the self and others. Resent research in cognitive science has added to what was understood in performance theory, claiming that our interaction with the world is not restricted to reacting to enacted events but is fundamentally built on understanding and prediction events through the perception of the intention of others in the environment. Based on these theories, I have argued that a framework for interaction performance must have at it's foundation the ability of all agents to perceive the intention of the other agents and therefore be able to project their own intention.

This paper has present, along with the underling theories, a series of studies of systems developed around the principle of agents projecting their intention. The results of these studies indicate that it is possible to embed into response gestures of visual or audio media agents information that enables a performer to perceive the intention of the agent. However, further exploration and research in to the methods and metaphors is needed. A key issues that has been discovered by these studies is the significance of the cognitive load imposed on the performer from trying to predict the action of the system. Although there are indications that the manner in which the system expressed its intention does not need to be modeled on a known paradigm, these preliminary studies suggest that a stronger reference to signifiers that are already part of the performer's body knowledge reduces the performer's need to rationally analyze the intention of the system. However, the parameters around this issue have not been isolated. Still, a feeling of trust and sharing of space was experienced in the systems modeling the projecting of intention. These results are encouraging but more work will need to be done before such a system is to the point of collaborative interaction.

8. REFERENCES

[1] Bogart, A., and T. Landau. The Viewpoints Book: A Practical Guide to Viewpoints and Composition. New York: Theatre Communications Group, 2005

[2] Camurri, A., and P. Feffentino. "Interactive Environments for Music and Multimedia". Multimedia Systems 7: (1999): 32-47

[3] Camurri, A., et al. "The MEGA Project: Analysis and Synthesis of Multisensory Expressive Gesture in Performing Art Applications". Journal of New Music Research, 34:1 (2005): 5-21

[4] Gallese, V. "The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy". Journal of Consciousness Studies 8 5-7 (2001) 33-50

[5] Gallese, V., "The Intentional Attunement Hypothesis: The Mirror Neuron System and Its Role in Interpersonal Relations". Biomimetic Neural Learning for intelligent Robots. Berlin: Springer (2005) 19-30

[6] Iacoboni, M. et. al., "Grasping the Intentions of Others with One's Own Mirror Neuron System" PLoS Biology 3:3, (2005) 529- 535

[7] Kohler, E., et al. "Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons". Science 297: (2002) 846-8.

[8] Lee, D. "How Movement is guided" www.perception-inaction.ed.ac.uk/PDF_s/ Howmovementisguided.pdf

[9] Lewis, G. E., "Interacting with Latter-Day Musical Automata". Contemporary Music Review 18:3 (1999): 99-112

[10] Lockford, L. and Pelias, R., "Bodily Poeticizing in Theatrical Improvisation: A Typology of Performative Knowledge". Theatre Topics 14.2 (2004) 431-43

[11] Merleau-Ponty, M. The Visible and the Invisible. Evanston: Northwestern University Press, 1968

[12] Merleau-Ponty, M. Phenomenology of Perception. London: Routledge, 1962

[13] Ng, K. "Sensing and Mapping for Interactive Performance." Organized Sound 7:2 (2002) 191-200

[14] Rowe, R. "Incrementally Improving Interactive Music Systems". Contemporary Music Review 13:2 (1996) 47-62

[15] Saltz, D. "Live Media: Interactive Technology and Theatre". Theatre Topics 11(2001) 107-30

[16] Thom, B., "Artificial Intelligence and Real-Time Interactive Improvisation". Proceedings of the Seventeenth Conference on Artificial Intelligence, Austin Texas, August (2000)

[17] Wegner, D. M., The Illusion of the Conscious Will. Cambridge MA. USA: MIT Press 2002

[18] Weinberg, G., and Driscoll, S., "The Perceptual Robotic Percussionist- New Developments in Form, Mechanics, Perception and Interaction Design*"* 2nd ACM/IEEE International Conference on Human-Robot Interaction Washington DC, USA, March 9-11 (2007)

[19] Winkler, T. "Making Motion Musical: Gesture Mapping Strategies for Interactive Computer Music." Proceedings of the International Computer Music Conference. 1995

A Novel Approach in Blind Estimation of MIMO Channels Using Bussgang Algorithm

Maryam Dehghani Estarki [1], Ebrahim Karami [2] and Mohsen Shiva [3]

[1] Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

[2] Centre for Wireless Communications, University of Oulu, P.O. Box 4500, FIN-90014, Oulu, Finland

[3] ECE Department, Faculty of Engineering, University of Tehran, Campus#2, P.O. Box 14395-515, Tehran, Iran

email: mdehghan@sfu.ca, Ebkarami@ee.oulu.fi and mshiva@ut.ac.ir

***Abstract*—Providing high capacity services in wireless communications systems, is of great interest to researchers. Many researchers have been interested in multiple-input-multiple-output (MIMO) wireless systems that guarantee achievement of such goal [4] and [8]. One of the great challenges regarding a MIMO channel is to find the channel characteristics by estimating its impulse response which is time variant. In this paper, a novel scheme using a blind algorithm is presented to estimate the response of a MIMO channel. To the best of our knowledge, it is the first time that the Bussgang algorithm is used to estimate the channel. This novel method profits from the basic idea of the Bussgang Deconvolution algorithm, which was first proposed by Bellini, [1], [7] and [9]. This algorithm is iterative, completely self-tuning, and has a practically low-complexity structure. Computer simulations are presented to confirm the robustness of the analysis and illustrate the acceptable performance of this new approach by considering its tracking behavior, BER, stability, as well as its low complexity.**

***Index Terms*—MIMO channel, Bussgang algorithm, cost function adaptivity, bit error rate (BER)**

I. Introduction

RECENTLY, many researchers have been interested in MIMO wireless systems that guarantee achievement of high capacity services in wireless communication systems [4]. Where the channel bandwidth and the total transmitted power are constrained, it is known that by forming a channel, using increased spatial dimensions, one can achieve extraordinarily large capacities [4] and [6]. The main issue on MIMO channel study, is characterizing the time variant channel, and determining the channel impulse response. Although the physical phenomena constituting the channel response are very complex, and often non-linear, their effects upon the transmitted signal can be quite accurately modeled by a linear, possibly time-variant, system. Many methods have been proposed to achieve this goal. The presented algorithms for channel estimation can be divided in several categories. In a wide range of these algorithms, a training sequence is used to provide more accurate channel estimates. But in a mobile communication channel, it is impractical to employ a training sequence of long duration [7]. Therefore, many attempts have been made towards channel estimation without the use of training sequences or with the help of a short training sequence to establish a first proper estimate, and then to continue the estimation process without training. This group of etimation methods, is called blind methods. One of the most well-known blind algorithms is the Bussgang algorithm [1], [7], and [9]. The Bussgang algorithm is a blind algorithm mainly used to design the equalizers.

In this paper, the conventional Bussgang idea, is modified and employed to estimate the channel behavior. In this approach, an iterative algorithm is developed to adapt the Bussgang parameters. Then the Minimum Mean Squared Error (MMSE) criterion is employed in the receiver to detect the transmitted vector. Computer simulations are presented to confirm the robustness of the analysis and illustrate the acceptable performance of this new approach. We assume that the system is single user. Therefore, the effect of Multiple Access Interference (MAI) is not considered.

The paper is organized as follows: in section 2, a review of the conventional Bussgang method is presented, and then the problem formulation is stated. Section 3 presents the novel blind Bussgang-based iterative estimation method, the receiver block diagram, and signal detection criterion. In section 4, we introduce the simulation results for channel estimation, and symbol detection and describe the advantages of this system. And finally, in section 5, the paper is concluded.

II. Problem Formulation

In this paper, we use lower and upper case boldface letters to denote column vectors and matrices, respectively. Matrices $\mathbf{I}$, and $\mathbf{0}$, stand for the identity matrix and all-zero matrix, respectively. Superscripts $(.)^*$, $(.)^T$ and $(.)^H$ are sequentially denoting the complex conjugate, the transpose, and the complex conjugate transpose (hermitian).

A. Conventional Bussgang Method

Modified Bussgang algorithms for blind channel equalization are based on the Bayesian iterative estimation of the source sequence [2] and [3]. These algorithms, introducing a kind of source adaptivity, were developed to reduce the computational complexity of the original Bussgang algorithm as well as to make it more flexible [2]. In the blind deconvolution, only the output signal is known (the system and the input signal, are unknown), and the objective is to find both the input signal and the system

transfer function. In all deconvolution problems, we apply the same scenario.

Note that the parameters adaptivity by including the cost function adaptivity, makes the algorithm 'more blind' [3]. In the Bussgang technique, a nonlinear zero memory estimator, $g(.)$, is used to operate on the input signal, $z(t)$, at the processor part of the system. Then, the difference between the estimator, $g(z)$, and signal, z, contributes to form the estimate of the next snapshot [3]. In the original Bellini's theory, the optimal estimator depends on the statistics of the noise and the source sequence [9].

B. System model

The MIMO system of interest, equipped with N_t transmitting antennas and N_r receiving antennas, can be modeled as a matrix FIR filter. The channel between each transmit and receive pair is modeled as a flat fading process. $\mathbf{H}$ represents the $N_r N_t$ channel response matrix. The channel is Rician, which is a more general case than the Rayleigh channel. The Rician probability density function is often characterized by the ratio of the direct component power, A_0^2, to the diffuse component power, $2\sigma_d^2$ in decibel, K (dB), as below,

$$K\,(\text{dB}) = 10\log_{10}\left[\frac{A_0^2}{2\sigma_d^{\,2}}\right]. \tag{1}$$

For $K = -\infty$ dB, there is no direct path, and Rician distribution becomes Rayleigh. For higher values of K, the Rician distribution becomes almost Gaussian [13] and [14].

Each transmitted symbol vector is denoted by $\mathbf{b}$, an $N_t \times 1$ column vector whose components are statistically independent Gaussian variables. $\mathbf{n}$ is a $N_r \times 1$ column vector, with zero-mean additive white Gaussian noise elements with the variance of σ^2. It is assumed that the channel matrix, $\mathbf{H}$, and the noise vector, $\mathbf{n}$, are time-variant. The received $N_r \times 1$ vector is denoted by $\mathbf{r}$. Each receive antenna observes a linear combination of all transmitted data sequences, each distorted by ISI, under the white Gaussian noise. Specifically, the discrete time observable signal, from the j th receive antenna (with $j = 1,..., N_r$) at time t, $r_t^{(j)}$, is given by

$$r_t^{(j)} = \sum_{i=1}^{N_t} h_t^{(i,j)} b_t^{(i)} + n_t^{(j)}, \tag{2}$$

where $h_t^{(i,j)}$ is the channel impulse response between the i th input and the j th output of the MIMO channel at time t. The complex baseband constellation point $b_t^{(i)}$ is transmitted by the i th transmitting antenna at time t, and $n_t^{(j)}$ is the complex noise sample at the j th receiver. In essence, there exist a total of $N_r N_t$ interfering and time-varying SI channels, $h_t^{(i,j)}$ (with $i = 1,..., N_t$ and $j = 1,..., N_r$). A clear way to represent the time variation of the MIMO channel coefficients, is to rewrite the input-output relationship of (2) in vector form, by collecting $r_t^{(j)}$ from all receiver antennas and $n_t^{(j)}$ from all channels, at time t, into the N_r -dimensional column vectors, $\mathbf{r}$ and $\mathbf{n}$,

$$\mathbf{r} = \mathbf{H}\mathbf{b} + \mathbf{n}. \tag{3}$$

respectively, and rewriting the channel taps, $h^{(i,j)}$, at time t, in an $N_r N_t$ rectangular matrix, $\mathbf{H}$, the relation between vectors will be as follows:

Here the goal is to estimate $\mathbf{H}$, while receiving $\mathbf{r}$ at N_r receive antennas. Signal to Noise Ratio (SNR) is known from the transmitted signal and the channel specifications. The SNR is set to be the same value for all the transmit antennas. By choosing the factor K (dB), and neglecting the co-channel interference, the SNR of each interfering channel, is

$$\text{SNR} = 10\log\frac{1+10^{K/10}}{\sigma^2}, \tag{4}$$

because the 2-PSK points are normalized to unit power [11].

Given that all $N_r N_t$ (direct and interfering) ISI channels are normalized to have equal energy of $1+10^{K/10}$, the optimum allocation of power among the N_t transmitters, without special shaping, is power balanced [10] and [11]. Although, different channel energies and transmit power allocation can be explored, but in this paper, all simulations are performed based on the above assumptions.

III. The Proposed Algorithm

A. Modified Bussgang estimation

As stated in section 2, Bussgang equalizers benefit from a zero memory nonlinear function which changes in order to approach the real value of the unobserved signal. This process is performed iteratively; therefore, the function is updated repeatedly and results in obtaining the modified version of the Bussgang equalizer, which is used to achieve a good estimate of the unknown channel and the unobserved transmitted sequence.

We define the cost function based on the modified system, and then by minimizing the cost function, the appropriate value of the channel matrix is obtained. Because the channel is time-varying, this process is done repeatedly. The new calculated $\mathbf{H}$, is the input of the signal detector. In this approach, we have exploited an MMSE detector. The estimated matrix, $\hat{\mathbf{H}}$, and the received signal, $\mathbf{r}$, are the inputs of the detector. In Fig. 1, the block diagram of the system of interest is depicted.

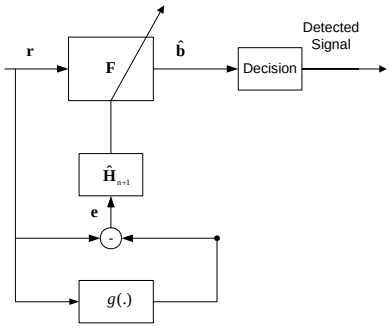


Fig. 1: Block diagram of the MIMO system receiver comprised of the Bussgang estimator and the MMSE sequence detector.

The received vector, $\mathbf{r}$, is fed to the MMSE detector which is shown by box $\mathbf{F}$. $\hat{\mathbf{H}}_{\mathbf{n+1}}$, denotes the new estimate of the channel matrix, $\mathbf{H}$, and $\mathbf{r}$ denotes the inputs of the MMSE algorithm. $\hat{\mathbf{b}}$, the output of box $\mathbf{F}$, is the soft decision of the transmitted signal, $\mathbf{b}$. Function $g(.)$, operates on $\mathbf{r}$. The difference between $\mathbf{r}$ and $g(\mathbf{r})$ forms the error signal, $\mathbf{e}$, which is the input of the estimator block to form the next estimate of the channel matrix. $\hat{\mathbf{b}}$ is fed to the decision block, which is a simple sign operator because of the 2-PSK modulation. This block can be designed according to the type of the modulation used in the system.

As stated, first we define the cost function, J, as below

$$\mathrm{J}(\hat{\mathbf{H}},\mathbf{r}) = (\mathbf{r} - g(\mathbf{r}))^2 . \quad (5)$$

In order to achieve the minimum value of the cost function, it suffices to compute its derivative. J is a quadratic function; thus, it has minima. By computing the derivative of $\mathrm{J}(\hat{\mathbf{H}},\mathbf{r})$, it follows

$$\Delta\hat{\mathbf{H}} = \eta[-(g'(\mathbf{r}) - \mathbf{1})(g(\mathbf{r}) - \mathbf{r})]\hat{\mathbf{b}}, \quad (6)$$

where η is a positive step size and $g'(\mathbf{r})$ denotes the derivative of the function $g(\mathbf{r})$ with respect to $\mathbf{r}$.

Some well known functions are used to form the Bussgang operator, $g(.)$. For a wide range of noise power, a suitable choice is the bilateral Sigmoid function as follows,

$$g(z) = c\ \tanh\ (dz), \quad (7)$$

while c and d are chosen properly [5] and [16]. Clearly a suitable pair of constant parameters, c and d, can not be fixed since all the conditions of the system, such as noise, interference, etc. are varying. In order to update c, we use the approach suggested in [7]:

$$c = \frac{\mathrm{E}[\mathbf{r}^*\mathbf{r}]}{\mathrm{E}[\mathbf{r}^* c \tanh d(\mathbf{r})]}. \quad (8)$$

To obtain d, we use the Gradient Steepest Descent (GSD) method,

$$\Delta d = \lambda c[(\mathbf{r} - c\tanh(d\mathbf{r}))^* (\mathrm{sech}(d\mathbf{r}))(\mathrm{sech}(d\mathbf{r}))^* \mathbf{r}] \quad (9)$$

where λ is the positive arbitrary step size to be adjusted accordingly. By substituting the new values in the cost function equation, the new estimate of $\hat{\mathbf{H}}$ is simply computed as below

$$\hat{\mathbf{H}}_{\mathbf{n+1}} = \hat{\mathbf{H}}_{\mathbf{n}} + \mu \cdot \Delta\hat{\mathbf{H}}. \quad (10)$$

For each vector of the received signal, this process is repeated. Therefore, the channel is continuously updated. For better channel tracking, and reaching a good starting point for the channel estimation, we can use a training sequence and pursue the estimation process by one of the known algorithms (such as LMS, ML, EM, etc.), and then continue the estimation process blindly. As well, we can allocate a very small portion of each transmitted block to training symbols, and then switch the receiver to decision-directed mode and continue the estimation and detection process with no pilot.

The simulation results show that a small percentage of pilot (training) symbols have a great influence on the performance of the system and decreases the BER significantly.

B. Signal detection using MMSE criterion

In order to generate estimates of the transmitted symbols, we use the MMSE criterion to adjust the tap weight coefficients [12]. $\hat{\mathbf{b}}$, the optimum MMSE estimate of the transmitted signal, $\mathbf{b}$, is generated based on the estimated channel response matrix, $\hat{\mathbf{H}}$, and the received vector, $\mathbf{r}$, This algorithm provides a computationally efficient means

$$\mathrm{F}(\mathbf{r}) = [\hat{\mathbf{H}}'\hat{\mathbf{H}} + \sigma^2\mathbf{I}_{\mathbf{N_t}}]^{-1}\hat{\mathbf{H}}'\mathbf{r}. \quad (11)$$

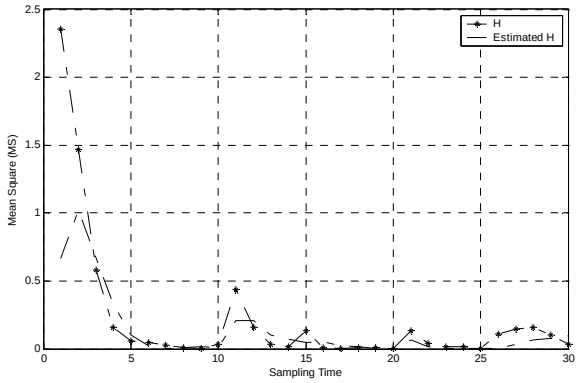
to accomplish the task of signal detection. The detection method described above is a standard solution.

IV. Simulation Results

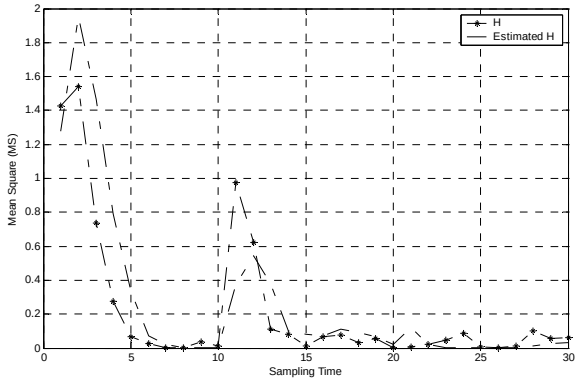
We have developed the above estimation and detection process for different structures of transmit and receive antennas in different channel conditions. The proposed algorithm has shown very good performance under each imposed condition. The channel estimation algorithm is also used for the case of semi blind, where a portion of each block of the transmitted signal is allocated to the training symbols. By changing the percentage of the pilot bits, the performance of this novel method is evaluated by considering the BER. Also, the tracking of the channel is experienced by changing the effective parameters.

A. Channel estimation

We consider the tracking behavior and the BER as performance criteria for simulated channels. In Fig. 2, the tracking of the channel is depicted for two different values of Rician Factor, K. The real random channel impulse response and the estimated impulse response are tracked for a large sampling period. Each unit on the horizontal axis consists of 10 blocks of 256 bits, along which the Mean Square (MS) error is computed. The tracking behavior of the channel is quite acceptable. We observe that the estimation has the same fluctuations as the real channel.



(a)



(b)

Fig. 2: MS of the real random channel and the estimated one via the suggested Bussgang scheme for two different channels: $N_t = 2$, $N_r = 6$, SNR=10 dB, (a) $K = 10$ dB, (b) $K = 20$ dB.

Although the channel has an impulsive change, the estimation tracks it with the same pace. The low difference between these two values, represents the small MSE of the estimated channel components.

B. Symbol detection

In Fig. 3, the performance versus SNR is depicted. Four transmit antennas are used in a channel of Rician factor K of 10dB. This process is repeated for different numbers of receive antennas.

Fig. 4 shows the performance versus the number of receive antennas with 4 transmit antennas of 20 dB Rician factor for different power allocations of the transmitted signal. For the same amount of power, the algorithm has better performance as the number of receive antennas increases. The process is repeated for different Rician channels where it is shown that for larger Rician factor (in dB) the performance improves.

In Fig. 5, performance of the proposed algorithm is shown versus the percentage of training symbols in each block of the transmitted signal. The algorithm is performed by 4 transmit and 6 receive antennas with 10 dB of SNR for different ratios of Rician factor. In Fig. 5, performance of the proposed algorithm is shown versus the percentage of training symbols in each block of the transmitted signal. The algorithm is performed by 4 transmit and 6 receive antennas with 10 dB of SNR for different ratios of Rician factor.

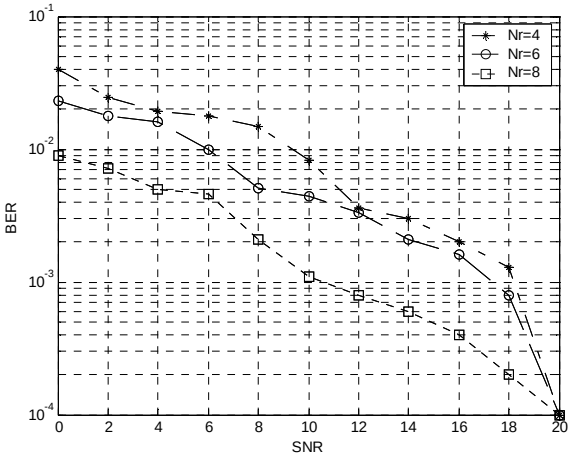


Fig. 3: Performance of the proposed algorithm for different SNR with $N_t = 4$ and 10 dB Rician factor.

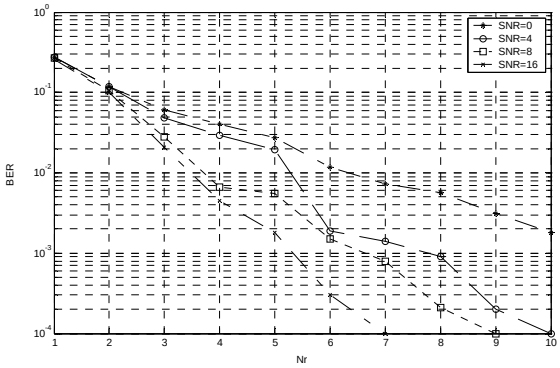


Fig. 4: Performance of the proposed algorithm for different N_r, with $N_t = 4$ and 20 dB Rician factor.

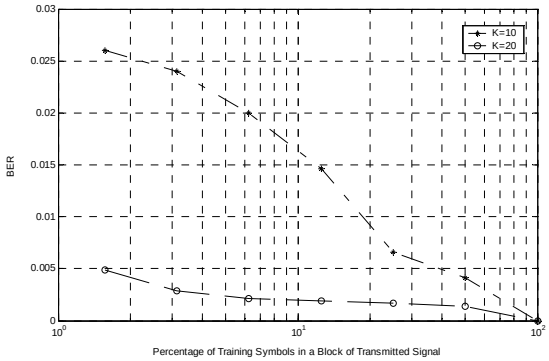


Fig. 5: Performance of the proposed algorithm for different training percentages with N_t=4, N_r=6 and SNR=10.

C. Advantages

The advantages of this novel approach when compared with the conventional methods are evident. Most of the blind methods are very complex and time consuming. Some of them need extra information about the system condition. The most important contribution of this method, is its simplicity and low complexity in the suggested practical system. The system has lower complexity, lower cost, and more possibility for realization over most of the proposed algorithms.

Another merit of this novel approach is its good performance in terms of the BER, and tracking behavior.

It can be implemented as a fully blind processor. As well, it is adjustable to be used as a pilot assisted one; thus, it can be a fully blind or a semi-blind algorithm. Therefore, under the harsh condition of the channel, or when the power of the transmitted signal is insufficient, the system can be switched to a more pilot-assisted mode. In other words, the degree of blindness of the estimator is adjustable.

V. Conclusion

In this paper, the problems of blind channel estimation and symbol detection are addressed for a typical MIMO system operating in frequency flat fading environments. A channel estimation scheme is proposed, which yields consistent channel estimates. The proposed scheme is simple to implement and capable of dealing with interference from various sources. Computer simulations are presented to confirm our analysis that illustrates the performance of this novel algorithm. In all the simulations presented in this paper, the receiver algorithm outlined above is implemented and compared to the performance of the system, by assuming different numbers of receive and transmit antennas in single user Rician channels with different amounts of power for the transmitting signal. The performance is also studied for the different Rician factors, K.

References

[1] S. Bellini, "Blind Equalization," *Alta Frequenza*, vol. 57, pp. 445-450, 1988.

[2] S. Fiori, "Analysis of Modified 'Bussgang' Algorithm (MBA) for Channel Equalization," *IEEE Transactions on Circuits and Systems*, Part I. Accepted for publication. http://www.unipg.it/~sfr/publications/TCASI04.pdf.

[3] S. Fiori, "Cost Function Adaptivity in Bussgang Filtering," *Electronics Letters*, Vol. 39, No. 6, pp. 572-573, March 2003

[4] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications* vol. 6, no. 3, pp. 331-335, March. 1998.

[5] G. H. Golub, and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Fifth printing 1987.

[6] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951-964, April 2003.

[7] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1996 (Chapter 18: Blind Deconvolution)

[8] W. C. Jakes, Jr., *Microwave Mobile Communications*, John Wiley and Sons, New York, 1974, Chapter 1 and 5.

[9] E. Karami and M. Shiva, "Adaptive beamforming algorithm for antenna array in multipath fading channels CDMA systems," in *Proceeding of 8th IEEE international symposium on communication and vehicular technology*, Delft, Netherlands, 18 Oct. 2001, pp. 108-112.

[10] C. Komninakis, "Joint Channel Estimation and Decoding for Wireless Channels," A dissertation submitted for the degree Doctor of Philosophy in Electrical Engineering, University of California, Los Angeles, 2000.

[11] C. Komninakis, C. Fragouli , A. Sayed, and R. D. Wesel, "Multi-input multi-output fading channel tracking and equalization using Kalman estimation," *IEEE Transactions on Signal Processing*, vol. 50, No. 5, pp.1065-1076, May 2002.

[12] J. G. Proakis, *Digital Communications*, Fourth Edition, 2001,McGraw Hill.

[13] P. M. Shankar, *Introduction to Wireless Systems* , 2002, John Wiley & Sons, Inc.

[14] Hazysztof Wesolowski, *Mobile Communication Systems*, 2002, John Wiley & Sons, Ltd.

Maryam Dehghani Estarki (M'98) was born in Kashan, Iran, on March 21, 1989. She received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tehran, Tehran, Iran, in 2000 and 2004, respectively. From 2006 she has been a Ph.D. student at the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. She is now a research assistant and working under the supervision of Professor Rodney Vaughan.

.

Iterative Probabilistic Distribution Learning Algorithm for Multi-user Space Time Block Coded Receiver

Muhammad Naeem
Simon Fraser University, School of Engineering Science
8888 University Drive, Burnaby, BC V5A 1S6 Canada
{mnaeem, dchlee }@sfu.ca

Abstract— In this paper, an Iterative Probabilistic Distribution Learning Algorithm (IPDLA) is proposed to detect symbols in Space-Time Block Coded (STBC) Multi Input Multi Output (MIMO) Communication System. We transform the STBC-MIMO system into a virtual multiuser STBC-MIMO system, where each user has multiple antennas. This transformation enables us to apply advance Multiuser Detection (MUD) techniques for detection. The STBC-MUD is known to be a NP-Hard problem. Using exhaustive search, the optimum detector or Maximum Likelihood (ML) has a computational complexity that exponentially increases with the number of users, transmit antennas, and complex modulation schemes.

The IPDLA not only searches for a near-optimum solution in real time but also has very low complexity as compared to the ML detector. The IPDLA finds the near optimum solution by learning probability distribution from previous iterations. We also proposed a two stage IPDLA, exploiting the non-converging probabilities of probability distribution and then apply Local Search on these bits (we call it IPDLA-LS). Monte Carlo simulations are used to investigate the performance of the proposed algorithm with Zero Forcing (ZF), MMSE, VBLAST, and ML detectors in terms of bit error rate (BER), convergence, and complexity.

Index Terms— space time block code, Multi-user detection, Probabilistic distribution learning

I. INTRODUCTION

RECENT research on wireless communication systems shows that using multiple antennas at the both the transmitter and receiver enables higher data rate communication as compared to the single antenna system [1]. The MIMO system has the ability to deal with multipath propagation. It effectively takes advantage of random fading and multipath delay spread [1][2]. It is also shown in [3] that the use of space-time block codes improves the capacity and unprecedented spectral efficiencies over wireless channels.

In this paper, our focus is on an uplink scenario. In [4] Verdú proposed the single antenna based ML MUD that offers the best BER performance among all other detectors. In [5] Poor first proposed a space-time MUD algorithm. Space-time coded MUD techniques for Code Division Multiple Access (CDMA) is investigated in [6]. One of the optimum solutions to Space-Time Block Coded (STBC) MUD is the Maximum Likelihood (ML) detector but it requires tremendous processing power at the receiver side. The ML detector uses exhaustive search over all possible user symbol sequences [7], so its computational complexity increases exponentially with number of users and modulation scheme. Due to the high complexity of the ML detector, a number of suboptimum solutions were proposed (e.g., Conventional Detector (CD), Zero Forcing (ZF), and Minimum Mean Square Error (MMSE) [4-5])

In CD, a bank of matched filters is used to detect each user while considering other users as noise. Due to the near-far effect, it suffers from Multiple Access Interference (MAI). The ZF detector attempts by nulling out the interference by direct inverting the channel with weight matrix. Hence, in the absence of channel noise it achieves perfect detection unlike CD [7], however, in the presence of channel noise, it leads to noise enhancement, which in turn leads to higher BER. Another class of linear detector is MMSE that minimizes the mean square error between the transmitted symbol and the decision variable. It also circumvents the noise enhancement problem. In near optimum detectors a Genetic Algorithm Detector (GAD) based STBC MUD detector was proposed in [8]. The problem with GAD is that it requires several parameter values to be fine tuned to achieve good results. Second in GAD it is difficult to predict the evolution of the population, and also, good blocks can be broken by the effect of crossover operators.

By keeping in mind all the problems of previously proposed detectors, in this paper we introduce an iterative probabilistic distribution-learning detector. The Iterative Probabilistic Distribution Learning Algorithm (IPDLA) has its roots from estimation of distribution algorithms (EDA) [9]. These algorithms are inspired by learning and simulation of Bayesian networks as the foundation of distribution estimation. The IPDLA MUD operates on simultaneous/parallel solutions. Each iteration results in a new set of possible solutions created by estimating the joint probability distribution of the previous iteration. No parameter setting is required for this algorithm as compared to GAD. We also proposed a modified IPDLA with Local search, which further improves the performance of the detector.

The rest of the paper is organized as follows. The system model and optimum symbol detection of STBC-MIMO-MUD is presented in section II. Section III proposes the IPDLA. Section IV provides the computational complexity comparison.

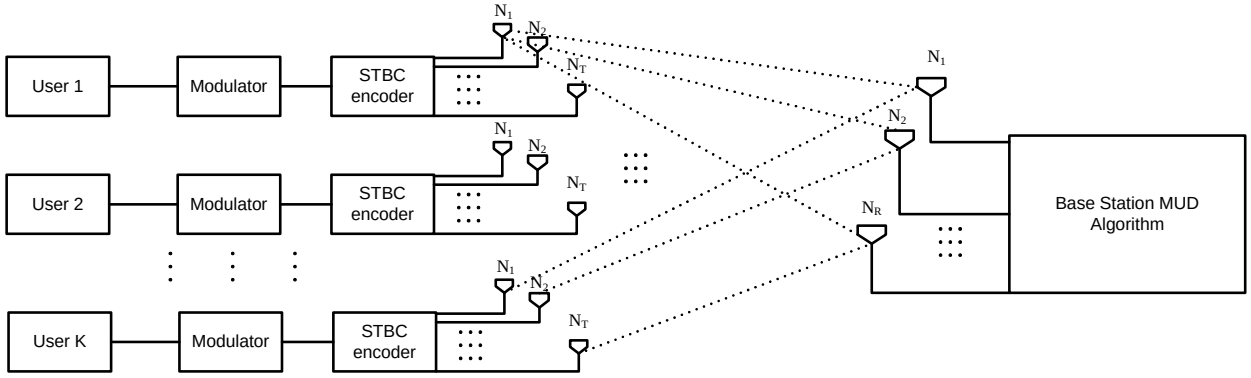


Fig. 1: A block diagram of Multiuser STBC uplink system.

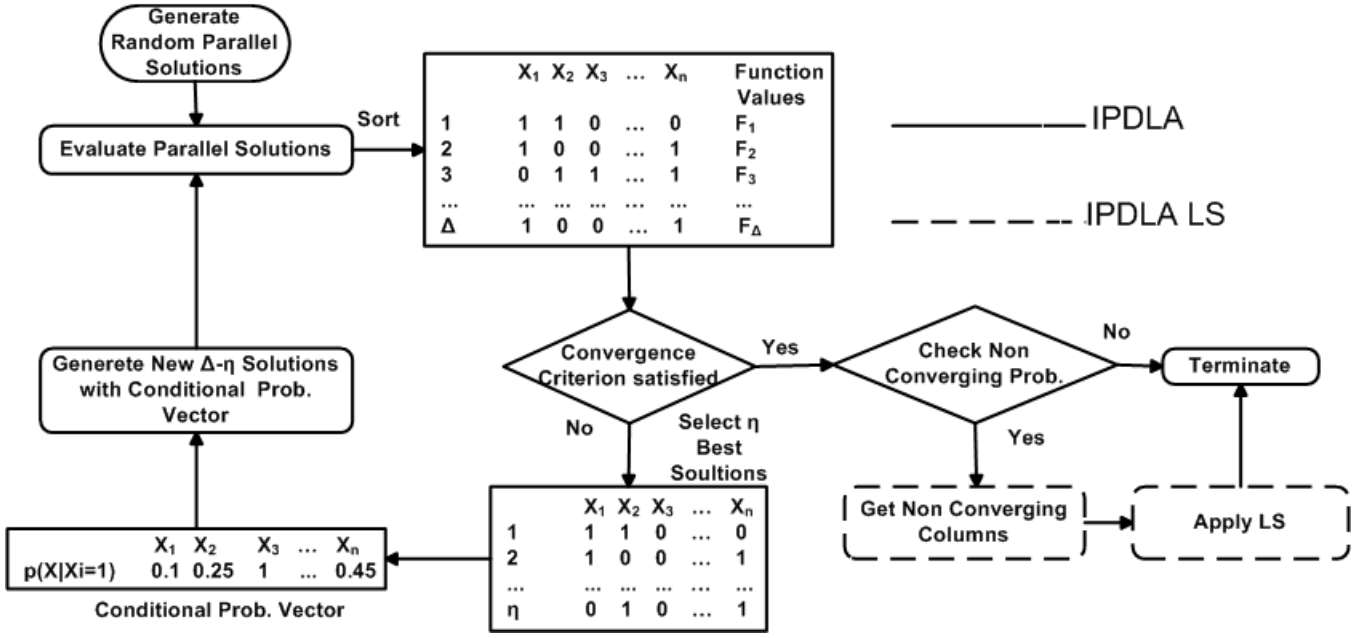


Fig. 2: Algorithm Flow Diagram

Simulation results are presented in section V. Finally section VI concludes this paper.

I. SYSTEM MODEL

We consider a spatial multiplexing MIMO multiuser system with *K* users, N_T transmit antenna for each user, and *R* receive antenna at the base station. Figure 1 shows a multiple antenna assisted multiuser space-time block coded uplink scenario. Source symbols are mapped into complex modulation schemes of M-PSK and M-QAM. The channel model is assumed to be quasi-static, and mean channel gain remains constant during one block of data. It is also assumed that channel state information (CSI) is perfectly known at the receiver and remains constant during an entire block length. This assumption is reasonable if training or pilot signals are sent to learn the channel, which is constant for some coherence interval. In this paper, rate 1 Alamouti scheme [9] is used for STBC. According to the Alamouti scheme, the transmitted symbols for k_{th} user is encoded as

$$S = \begin{bmatrix} s_1^k & s_2^k \\ -s_2^{*k} & s_1^{*k} \end{bmatrix} \quad (1)$$

where s_i^k represent the symbol transmitted by k_{th} user. Rows in (1) represent the time slots, and columns shows the symbols from different antennas. A generalized expression for *K* users, received signal at the r_{th} receive antenna, at time slot *t=1* will be

$$y_r = \sum_{k=1}^{K} \sum_{n=1}^{N_T} s_n^k h_n^{k,r} + z_r \quad (2)$$

For this paper we are using $N_T = 2$ per user .In Vector form, Multiuser space time coded received signal at the r_{th} receive antenna, (for a pair of time slots i.e., $t=1$ and $t=2$) is

$$\begin{bmatrix} y_{r,1} \\ y_{r,2} \end{bmatrix} = \begin{bmatrix} s_1^1 & s_2^1 & \cdots & s_1^K & s_2^K \\ -s_2^{*1} & s_1^{*1} & \cdots & -s_2^{*K} & s_1^{*K} \end{bmatrix} \begin{bmatrix} h_1^{1,r} \\ h_2^{1,r} \\ \vdots \\ h_1^{K,r} \\ h_2^{K,r} \end{bmatrix} + \begin{bmatrix} z_{r,1} \\ z_{r,2} \end{bmatrix} \quad (3)$$

where $y_{r,t}$ represents receive data at r_{th} antenna on t_{th} time slot. s_i^k is the k_{th} user's i_{th} data symbol and $h_N^{k,r}$ is the channel gain of k_{th} user between n_{th} transmit antenna and r_{th} receive antenna. We can also write (2) as

$$\begin{bmatrix} y_{r,1} & y_{r,2}^* \end{bmatrix} = \begin{bmatrix} s_1^1 & s_2^1 & \cdots & s_1^K & s_2^K \end{bmatrix} \begin{bmatrix} h_1^{1,r} & h_2^{*1,r} \\ h_2^{1,r} & -h_1^{*1,r} \\ \vdots & \vdots \\ h_1^{K,r} & h_2^{*K,r} \\ h_2^{K,r} & -h_1^{*K,r} \end{bmatrix} + \begin{bmatrix} z_{r,1} & z_{r,2}^* \end{bmatrix} \quad (4)$$

For R receive antennas

$$\begin{bmatrix} y_{1,1} & y_{1,2}^* & \cdots & y_{R,1} & y_{R,2}^* \end{bmatrix} = \begin{bmatrix} s_1^1 & s_2^1 & \cdots & s_1^K & s_2^K \end{bmatrix} \begin{bmatrix} h_1^{1,1} & h_2^{*1,1} & \cdots & h_1^{1,R} & h_2^{*1,R} \\ h_2^{1,1} & -h_1^{*1,1} & \cdots & h_2^{1,R} & -h_1^{*1,R} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ h_1^{K,1} & h_2^{*K,1} & \cdots & h_1^{K,R} & h_2^{*K,R} \\ h_2^{K,1} & -h_1^{*K,1} & \cdots & h_2^{K,R} & -h_1^{*K,R} \end{bmatrix} + \begin{bmatrix} z_{1,1} & z_{1,2}^* & \cdots & z_{R,1} & z_{R,2}^* \end{bmatrix} \quad (5)$$

Equation (5) can be written as

$$Y = S H + Z \quad (6)$$

where $Y \in \mathbb{C}^{1\times 2R}$ is the received vector, $S \in \mathbb{C}^{1\times 2K}$ is the transmitted symbol vector, channel matrix $H \in \mathbb{C}^{2K\times 2R}$, and additive noise vector $Z \in \mathbb{C}^{1\times 2R}$.

Transmitted symbols from a known finite alphabet $S \in \{s_1, s_2, ..., s_M\}$ of size, M, for each user at n_{th} transmitted antenna are passed to the channel. The detector chooses one of the possible combinations of $M^{N_T K}$ possible transmitted symbols. Assuming that the symbol vectors $S \in \mathbb{C}^{1\times 2K}$ are equiprobable, the Maximum Likelihood (ML) detector always returns an optimal solution according to the following:

$$\hat{S} \triangleq \arg\max_{S \in \mathbb{C}^{1\times 2K}} P(Y \mid S) \quad (7)$$

Assuming the additive noise, $Z \in \mathbb{C}^{1\times 2R}$, to be white and complex Gaussian, the ML detection can be expressed as the minimization of Euclidean distance to a target vector ,Y, over a $N_T K$ dimensional finite discrete set or $N_T K \log_2(M)$ dimensional finite binary set:

$$\hat{S} \triangleq \arg\min_{S \in \mathbb{C}^{1\times 2K}} \|Y - SH\|^2 \quad (8)$$

The optimal ML detection scheme needs to search all $M^{N_T K}$ or $2^{bN_T K}$ symbol combinations (b is the number of bits per symbol). Enumerating over all possible combinations, and finding the one that can minimize the (8), is an NP-Hard problem. The computational complexity increases exponentially with constellation size, M, N_T transmit antennas, and K number of users. High speed processing requirements of real time applications demand a comparatively simplified, fast and low complexity detection scheme

One approach to do detection is the Zero Forcing (ZF) Detector which require the inverse of channel matrix

$$\hat{S} \triangleq H^{\dagger} Y \quad (9)$$

where $H^{\dagger}$ denotes the Moore-Penrose pseudoinverse of H, which is a generalized inverse that exists even when H is rank deficient [11].A major drawback of ZF is that nulling out the interference without considering the noise can boost up the noise power significantly [12], which in turns results in performance degradation especially at low SNR.

Another technique is MMSE that minimizes the mean squared-error [13]. MMSE includes the noise power in filter tap calculations. MMSE controls noise boost up by attenuating all the users to noise level.

$$\hat{S} \triangleq \left(H^H H + \sigma^2 I\right)^{-1} H^H Y \quad (10)$$

where σ^2 is the noise power. Although MMSE is better than ZF however it is still not near to the optimal detector.

II. Iterative Probabilistic Distribution Learning Algorithm

Generally, two techniques are used to solve NP-Hard problems. One is the branch and bound (B&B) which require high computational complexity time, so it is not recommended for large size or time sensitive problems. Another approach is the Evolutionary Algorithm (EA) that searches for near optimum solutions to obtain good results within a reasonable computing time. The problem with EA is that it depends to a large extent on associated parameters like operators, probabilities of crossover, probabilities of mutation, and replacement strategies.

IPDLA is a population based search algorithm [10] that relies on probabilistic modeling of potential solutions in combination with the simulation of induced models to guide

their search. In this algorithm, the new population individuals are generated without crossover and mutation operators; instead, probability distribution is estimated through the joint probability distribution associated with the individuals selected at each iteration from the pool containing only the best individuals from the previous generations. A generalized IPDLA can be defined as $IPDLA = (I, F, \Delta, \eta, p_s, \Psi, \Upsilon)$, where

1. I is the space of individuals (a set of simultaneous parallel solutions). In the problem in hand the search space is $I = \{0,1\}^n$ where $n = N_t K \log_2(M)$.
2. F denotes a fitness function.
3. Δ is the maximum number of simultaneous/parallel solutions at a single iteration.
4. η is the number of best solutions from Δ solutions.
5. p_s is the selection probability to select η solutions.
6. Ψ is the estimated distribution from η solutions.
7. Υ is the termination criteria.

The basic IPDLA algorithm is illustrated in figure 2 and 3. Each solution of IPDLA has length $n = N_t K \log_2(M)$ bits. The initial solutions are obtained by sampling of probability distribution

$$p(X^j) = \prod_{i=1}^{n} p(x_i^j) \; \forall j = \{1, 2, ..., \Delta\} \tag{11}$$

The initial distribution is uniform, so $\forall i = \{1, 2, ..., n\} : p(x_i = 1) = p(x_i = 0) = 0.5$. The solution of length, *n*, is a vector of random variable, $X^j = (x_1^j, x_2^j, x_3^j, ..., x_n^j)$, where $x_i^j \in (0,1)$. The current iterative solution can be written as

$$(X^1, X^2, X^3, ..., X^\Delta) = \begin{Bmatrix} (x_1^1, x_2^1, x_3^1, ..., x_n^1), \\ (x_1^2, x_2^2, x_3^j, ..., x_n^j), ..., (x_1^\Delta, x_2^\Delta, x_3^\Delta, ..., x_n^\Delta) \end{Bmatrix} \tag{12}$$

In each iteration, η best solutions are selected from current Δ solutions. This selection can be accomplished using following expression

$$\Psi = \prod_{1}^{n} P(x_i \mid I_{t-1}^{\eta}) \tag{13}$$

The basic principle of IPDLA can be illustrated using the following example, a well-known binary optimization problem known as one-max function, $F = \sum_{i=1}^{n} x_i$, $\forall x_i \in (0,1)$.

Figure 3 shows four blocks of basic IPDLA algorithm [14]. In the given example, the task is to maximize the fitness function. The number of parallel solutions (Δ) are 10 and *n=6* and $p_s = 0.5$. For initialization we consider a uniform distribution $\forall i = \{1, 2, ..., 6\} : p(x_i = 1) = p(x_i = 0) = 0.5$ .After evaluating the fitness function, we choose $(p_s \times \Delta) = 5$ best solutions. In the next step we estimate the probability distribution using (13) and then we generate new solutions according to the estimated probability distribution. In the last step, we replace the worst solutions with these newly generated solutions. The process is repeated till the convergence criterion is met. The same procedure is applied to STBC-MUD problem. Here the fitness function is to minimize (8), we convert the problem to a binary optimization problem by searching $M^{N_T K}$ symbols to $2^{bN_T K}$ bits search. Each of the solutions has length, $n = N_T K \log_2(M)$ bits.

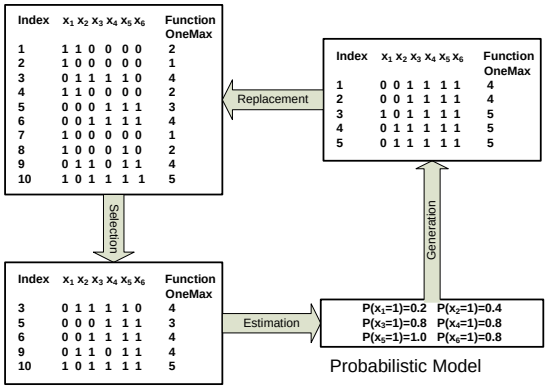


Fig. 3: The basic understanding of iterative probability generation

We also introduce an improved version of the IPDLA algorithm. The experimental results show that some of the probability distribution does not converge to 1. So we set a threshold μ for the probability distributions to converge. If the values of any of the probability distributions is less than μ, we apply Local Search to these bits.

III. Computational Complexity

The main purpose for the proposed near optimal algorithm is to illustrate their low computational complexity, in this section, we enumerate the computational complexity of ZF, MMSE, ML, IPDLA and IPDLA local search methods for STBC-MIMO-MUD problem. The complexity is counted as a time bound for algorithm , we will use $O(\bullet)$ notation as a metric to show the time bound. It is well known that the computational complexity of ML detector is $O(2^n)$ where $n = N_T K \log_2(M)$ [7]. It means we must evaluate all possible solutions and select the one that provides the minimum. However, this exhaustive search is impossible from a practical point of view. We need some sub-optimal detector. The ZF is one of the sub-optimum detectors. Since there is an inverse of a complex matrix is involve for ZF, the computational complexity of inverse of complex matrix is $O(n^3)$. The MMSE detector has the same computational complexity as of ZF detector.

The computational complexity of IPDLA is *approximately* $O(\eta I_t)$ if $\eta < \Delta$, where I_t is the number of iterations.(we use the word approximately because in the initialization phase, the algorithm needs to calculate Δ evaluations of fitness function). The worst case is if $\eta = \Delta$, then computational complexity is $O(\Delta I_t)$. Additional operation required by IPDLA is to sort,

generate a probability distribution and random number generation. For each iteration, we need to generate $\eta \times n$ random numbers. The computational complexity of random number generation is $O(1)$. The sorting computational complexity is $O(\Delta \log \Delta I_t)$, so the computation complexity for large Δ is $O(\eta I_t + \Delta \log \Delta I_t) \approx O(\Delta \log \Delta I_t)$.

The computational complexity of IPDLA-LS depends on the number of non-converging probability distribution bits. First we need to set a threshold for local search (e.g. n_{th}). Let the n_1 are the number of non-converging probability distribution bits. Then the computational complexity of IPDLA-LS will be $O(\Delta \log \Delta I_t + 2^{n_1})$ and for the worst case it will be $O(\Delta \log \Delta I_t + 2^{n_{th}})$. Normally n_{th} has a small value to reduce the computational complexity (e.g. $n_{th} \le 8$).In conclusion, IPDLA and IPDLA-LS offers very acceptable levels of complexity as compared to other detectors. Following parameters are used in our numerical results. The total number of service classes *M* = 4 with service rates equal to *R, 2R, 4R* and 8*R*, where *R* is the base transmission

IV. Simulation Results

For performance comparison, we present the simulation results of the proposed IPDLA and IPDLA-LS detectors with some of the existing detection techniques for STBC spatial multiplexing system. The channel is assumed to be quasi-static for each pair of time slots, but independent among different users. A summary of various simulation parameters are

- Users , $K=\{4,5,6\}$
- Number of transmit antennas per user $N_T = 2$
- Number Time slots $T_s = 2$ (Alamouti Scheme)
- Maximum number of simultaneous/parallel solutions at first iteration, $\Delta = \{50, 70, 100\}$ for $K=\{4,5,6\}$, respectively.
- Δ initialization : Random (With uniform probability)
- Number of best solutions to determine probability distribution is $\eta = \{25, 35, 50\}$ for $K=\{4,5,6\}$, respectively.
- Number of Iterations $I_t = \{30, 45, 50\}$ for $K=\{4,5,6\}$, respectively.

Figure 4 shows the BER performance comparison of ZF, MMSE, IPDLA and IPDLA-LS algorithms. The time to simulate the ML detector is huge, so we have only calculated the ML detector for the IPDLA and IPDLA-LS convergence results (see figure 8 and 9). The numbers of users are 4 with 4-QAM modulated data. The number of transmit antenna per users N_T are *2*. With selection probability of 0.5, the numbers of parallel solutions are 50 at the initialization. As shown by the simulation result, there is gain of 1.8 and 2.0 dB at BER of 10^{-2} for IPDLA and IPDLA-LS, respectively, over ZF and MMSE. Figure 5 and 6 show the BER performance for k=5 and k=6. Since by using more users, the computational complexity increases exponentially, we need to increase the values of Δ and the number of iterations.

The increase in the number of iterations adds complexity while the increase in the number of parallel solutions requires more memory. The tradeoff between parallel solutions and Number of iterations is illustrated in figure 7, which shows the required combination of parallel solutions and the number of iterations for any desired BER.

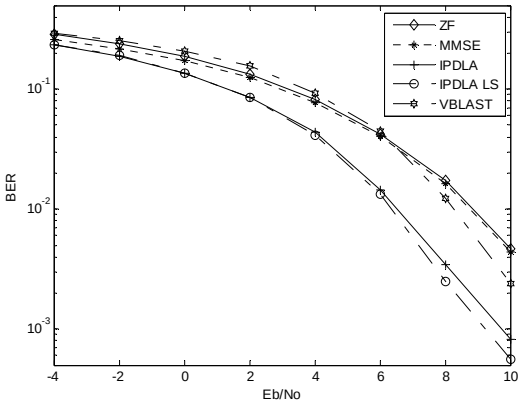


Fig. 4: BER performance for N_T=*2, K=4,* 4-QAM with N_R = *4* antenna configuration over quasi-static fading channel. The number of simultaneous/parallel solutions $\Delta = 50$, with $\eta = 25$ and $p_s = 0.5$. The number of Iterations are 30, Local search threshold is $n_{th} = 8$ (ML search space is $O(2^{16})$.

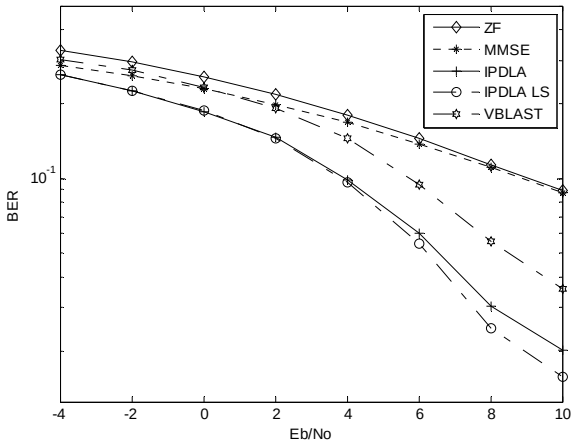


Fig. 5: BER performance for N_T=*2, K=5,* 4-QAM with N_R = *4* antenna configuration over quasi-static fading channel. The number of simultaneous/parallel solutions $\Delta = 70$, with $\eta = 35$ and $p_s = 0.5$ .The number of Iterations are 45, Local search threshold is $n_{th} = 8$ (ML search space is $O(2^{20})$.

Figure 8 shows the convergence of IPDLA and IPDLA-LS to the ML detector. The SNR is fixed to 6dB, and the numbers of parallel solutions are fixed to 30. The rest of the parameters are the same as that of figure 4. The only variable here is the number of iterations. We can see a rapid convergence of these algorithms to the ML, which shows the robustness of the proposed algorithm.

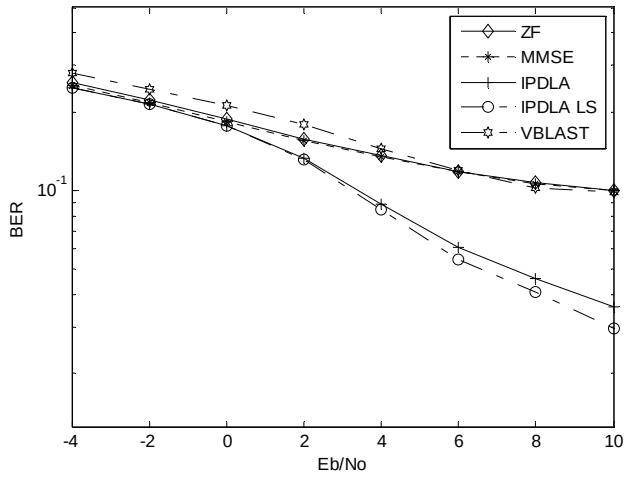


Fig. 6: BER performance for N_T=2, K=6, 4-QAM with N_R = 4 antenna configuration over quasi-static fading channel. The number of simultaneous/parallel solutions $\Delta = 100$, with $\eta = 50$ and $p_s = 0.5$ .The number of Iterations are 50, Local search threshold is $n_{th} = 8$ (ML search space is $O(2^{24})$.

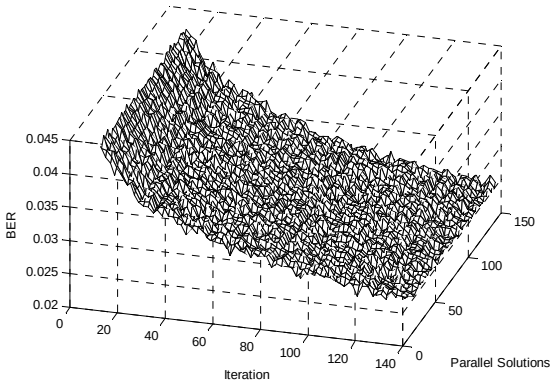


Fig. 7: Tradeoff between number of parallel solutions and Number of Iterations

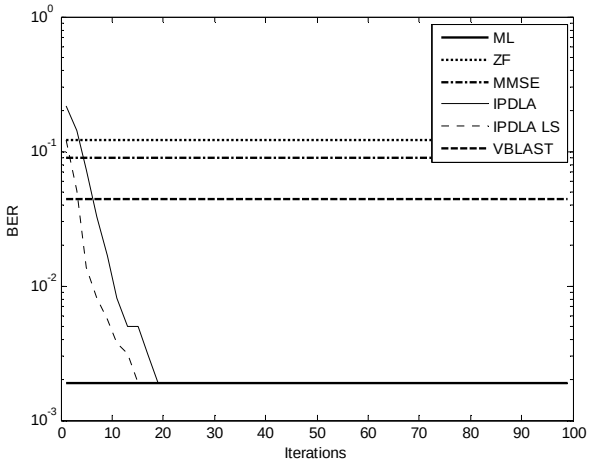


Fig. 8: Convergence of IPDLA and IPDLA-LS for fixed number of simultaneous/parallel solutions $\Delta = 30$, $\eta = 15$ and $p_s = 0.5$, N_T=2, K=4,SNR=6dB, 4-QAM with N_R = 4 antenna configuration over quasi-static fading channel. The number of Iterations are varied from 1 to 100, Local search threshold is $n_{th} = 8$.

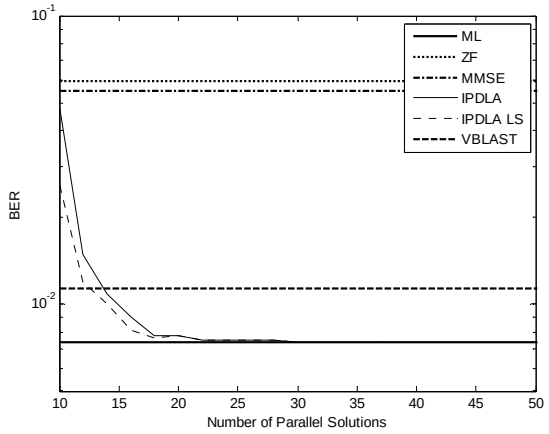


Fig. 9: Convergence of IPDLA and IPDLA-LS for fixed number of iterations $I_t = 30$, $p_s = 0.5$, N_T=2, K=4,SNR=6dB, 4-QAM with N_R = 4 antenna configuration over quasi-static fading channel. The number of parallel solutions are varied from 10 to 50, the local search threshold is $n_{th} = 8$.

V. CONCLUSION

In this paper, we proposed two probabilistic distribution learning algorithms, (IPDLA and IPDLA-LS) for Multi antenna base Multiuser Space Time coded wireless communication system. These detectors are simple in implementation and approach to near optimum performance. The complexity of these algorithms is very low as compared with the optimum detector. This simplistic model, lesser implementation complexity, resistance to trap in local minima and convergence to near optimum solution in less iterations , makes it suitable candidate for complex NP-Hard communication problems.

REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311-335, 1998.

[2] E. Telatar, "Capacity of multiantenna Gaussian channels," AT&T Bell Laboratories, Tech. Memo., June 1995.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communications: performance criterion and code construction," *IEEE Trans. Information Theory*, vol. 44, pp. 744-765, Mar. 1998.

[4] S.Verdú, "Minimum Probability of Error for Asynchronous Gaussian Multiple Access Channels", IEEE Trans. on Info. Theory, Vol 32, pp. 85-96, Jan 1986.

[5] X.Wang, V.H.Poor. "Space-time multiuser detection in multipath CDMA channels", IEEE transactions on signal processing, vol. 47, pp. 2356-2374, Sept. 1999

[6] H. Haung,H. Vishwanathan and G. J. Foschini, "Multiple antenna in cellular CDMA systems, Transmission, detection and spectral efficiency", IEEE Transactions on Wireless Communications, Vol.1, No.3, pp-383-392, July 2002

[7] S. Verdú, “Computational Complexity of Optimum Multiuser Detection”, Algorithmica, Vol.4, No.3, pp. 303-312, 1989.

[8] Y. Du and K .T. Chan, ”Improved Multiuser Detector Employing Genetic Algorithm in a Space-Time Block Coded System”, EURASIP J. of Applied Signal Processing, pp. 640-648, 2004

[9] S. M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications" ,IEEE journal on select areas in communications, Vol. 16, No. 8, pp 1451-1458 October 1998.

[10] Larrañaga P. and Lozano J. A. (2001) *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2001.

[11] G.H.Golub and C.F. VanLoan, *Matrix Computations*, The Johns Hopkins Univ. Press, 1983.

[12] S. Haykin, *Adaptive Filters*, 4th edition, McGraw Hill, NY,USA

[13] J. G. Proakis, *Digital Communications*, , 4th edition McGraw-Hill, NY, USA.

[14] E. Bengoetxea. “Inexact Graph Matching Using Estimation of Distribution Algorithms”. *PhD Thesis*. (2003) Brno university of technology.

Pedestrian Detection: A Novel Approach Using Segment Features

Mani Ranjbar

***Abstract*— This paper addresses the problem of pedestrian detection in a still image. Similar to previous feature-based methods, the detection process in the proposed method is accomplished by calculating the similarity between features of the input image and the pedestrian-specific learned features.**

The contribution of the proposed method is in presenting a novel feature called *segment feature*. A segment feature consists of the position and shape of a region in an image. To achieve a high performance result, the Adaboost algorithm is employed to generate a strong classifier, based on the weak classifiers created for each segment feature.

The performance of the proposed method is demonstrated using experimental results on INRIA datasets. Moreover, comparison results shows that the proposed method outperforms the detection ratio of the state-of-the-art approaches.

***Index Terms*—Pedestrian Detection, Segment Feature, Object Detection, Detection by Learning.**

I. Introduction

FINDING pedestrians in a scene is a challenging problem for several reasons. First, according to the application, a wide range of acceptable pedestrians' poses can be considered. Second, occlusion could add to the complexity of the problem. Third, the complexity of the scene background could impact the detection procedure (Figure 1 shows how this problem becomes more complex to solve due to the above issues). As depicted in Figure 1.d, in a real scenario, objects that may look like a pedestrian make this problem harder to solve.

The proposed method aims to create a bounding box over each pedestrian instead of finding the exact border. An example of a sample input image and the expected output image is shown in Figure 2.

In the proposed approach, image segments are considered as mid-level features and are called *segment features*. Each segment feature captures a connected region of the input image that has similar pixels inside. Utilizing these segment features, a large number of weak classifiers is defined. Each of the weak classifiers employs only one of this segment features as the classification measure. The Adaboost algorithm is then used for exploiting the overall classification capability of the weak classifiers.

Manuscript received February 28, 2008. This work was supported in part by the Simon Fraser University under the Pacific Century Scholarship.

M. Ranjbar is with the School of Computing Science, Simon Fraser University, British Columbia, Canada (phone: 604-725-6624, email: mra33@sfu.ca).

a b c d

Fig. 1. Different situations in pedestrian detection, a) simple case, b) complicated background, c) occlusion, d) similar non-pedestrian objects.

a b

Fig. 2. Sample pedestrian detection scenario, a) input image, b) output image.

The structure of this paper is as follows. In Section II, the previous work in the pedestrian detection literature is discussed. The proposed method is presented in detail in Section III. In Section IV, the experimental results of the proposed method are illustrated. In Section V, the conclusion of this paper is presented.

II. Previous Work

Different types of approaches have been considered to solve the problem of pedestrian detection. Particularly, in [1] some Harr wavelets are utilized as the input descriptors and rectified using a linear Support Vector Machine (SVM) to generate a

final classifier. Reference [2] employs a histogram of gradients (HOG) as its features and uses SVM to train the parameters of such classifiers. Using the HOG-based features, [2] achieves notable improvements compared to the results of [1]. Reference [2] also provides a large training and test dataset, INRIA, which contains a broad range of people posing under differing lighting conditions. Payam and Greg create *shapelet features* through a two step Adaboost algorithm [3]. They start with the image gradients in different orientations and combine them to make the shapelet features. The Shapelet approach significantly improves the performance of [2] in the INRIA dataset. In [4] some hypotheses are proposed using a local feature extractor and then a segment mask for each of these hypotheses is learned using a training pedestrian mask. This hypothesis-based method performs very well in the crowded scenes (including occlusion).

Other work is also reported in the literature for pedestrian detection in video sequences. For instance, Wren et al. applies background subtraction to find the moving objects to reduce clutter [5]. Zhao et al. report an approach in which the Markov chain Monte Carlo method is utilized for moving pedestrian detection [6]. In [7], occlusion is handled by training a color model for each person before occlusion, which is employed later to recognize occluded pedestrians.

III. Proposed Method

Briefly, the pedestrian detection task in the proposed method is performed as follows. The input image is scanned in different locations and scales. In each location and scale, a fixed size window (64×128 in the current implementation) is cropped from the input image and checked for the presence or absence of a pedestrian utilizing the following process:

1) Segment features of the window are extracted (see segmentation for detail),
2) The strong classifier, which is created in the learning process by combining the best pedestrian-specific features, is exploited to classify the window using its extracted features.

The following subsections describe each step of the proposed method in detail.

A. Segment Feature Definition

The segment feature is defined as a structure containing the shape and the location of a region in an image. More precisely, segment feature k is a 64×128 binary matrix which is one inside and on the border of the k^{th} region of the image and zero elsewhere. Consequently, each segment feature is a mid-level feature of the image because it contains a mid-level feature that is the region shape. Figure 3 shows some segment features of an image.

B. Segmentation Process

As mentioned above, the segment feature is generated by capturing the shape and the location of an image region. Therefore, a segmentation process is required to divide the input image into different regions (segments).

Among the proposed segmentation methods in the literature, the method proposed in [9] is employed because of its desired properties:

- Preserving the details in smooth areas while ignoring the details in textured regions,
- Adjusting the segmentation scale via the input parameter k,
- Accepting the color and gray scale images.

The following steps describe the segmentation process:

1) Each image pixel is considered as a region where it corresponds to a node (v in V) in the overall image graph of $G(V, E)$.
2) Neighboring pixels are connected by undirected edges (e in E). For each edge, a weight coefficient is computed according to dissimilarities among pixels.
3) Similar regions (A and B) are merged together to form a larger region if the following condition is held:

$$Dif(A,B) \le MInt(A,B), \tag{1}$$

where

$$Dif(A,B) = \min_{v_i \in A, v_j \in B, (v_i, v_j) \in E} w((v_i, v_j)), \tag{2}$$

where E is the graph edge set and for each node, and $w((v_i,v_j))$ is the weight between vertices v_i and v_j. Moreover,

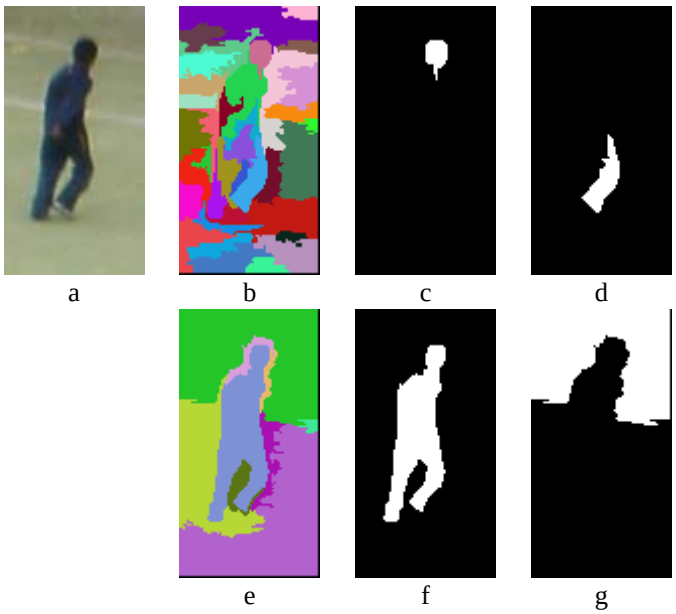


Fig. 3. Illustration of some segment features, a) input image, b) fine segmentation result, c,d) two segment features of fine segments corresponding to head and leg, e) coarse segmentation result, f,g) two segment features of coarse segments corresponding to the whole body.

$$MInt(A,B) = \min(Int(A) + \tau(A), Int(B) + \tau(B)) \quad (3)$$

$$Int(A) = \max_{e \in MST(A,E)} w(e) \quad (4)$$

$$\tau(A) = \frac{k}{|A|} \quad (5)$$

MST represents the *minimum spanning tree* of graph *G(V,E)*. The value of *k* defines the sensitivity level of the segmentation algorithm (i.e., larger *k* is preferred for larger regions). Note, however, that *k* is not the minimum segment size; smaller segments are allowed when there is a sufficiently large difference among neighboring regions.

In the pedestrian detection task, choosing a large *k* tends to segment the whole pedestrian body as one segment, but could generate incorrect segments when the pedestrian clothing is similar to the background. On the other hand, small values of *k* tend to capture different limbs, but at the cost of over segmenting (i.e., creating many segments for large limbs).

To capture the overall body shape as well as the shape of each limb, the input image is segmented with a large and a small *k* values (coarse and fine segmentation). Coarse and fine segments are used to generate the image segment features.

C. Weak Classifier Creation

After finding all the segment features in the training set, a weak classifier is generated using each segment features, which is defined as:

$$h_i(j) = \begin{cases} 1 & \Theta_l(i) < F_i(j) < \Theta_h(i) \text{ and } p_i = 1 \\ & \text{or} \\ & \Theta_l(i) > F_i(j) \text{ or } F_i(j) > \Theta_h(i) \text{ and } p_i = 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $F_i(j), \theta_l, \theta_h, p$ represent the value of feature *i* in training image *j*, low threshold, high threshold, and polarity value, respectively. Figure 4, depicts the process of creating feature vector F_i. θ_l, θ_j and *p* for each weak classifier are calculated using the following process:

1) Feature vector F_i is sorted in the ascending order,
2) The subsequence of F_i which produces the best classification result (contains the maximum ratio of positive or negative images) is found,
3) The value of F_i at the beginning and the end of this sequence is assigned to θ_l and θ_h , respectively,
4) *p* is assigned to one, if the subsequence contains the maximum ratio of positive images and to zero, otherwise.

D. Boosting Process

Because the best weak classifier can categorize only 76% of the training set accurately, the Adaboost method is employed to combine the best weak classifiers and create a strong classifier [8]. Unlike each weak classifier that may contain either the shape of one limb or the overall body shape, the strong classifier includes the combination of different limbs as well as the overall body shape. The strong classifier is defined as:

$$H(j) = \begin{cases} 1 & \sum_{i=1}^{T} [\alpha_i h_i(j)] \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where, *T* represents the number of weak classifiers used for generating the strong classifier, and α_i is the weight of weak classifier *i* obtained from the Adaboost algorithm.

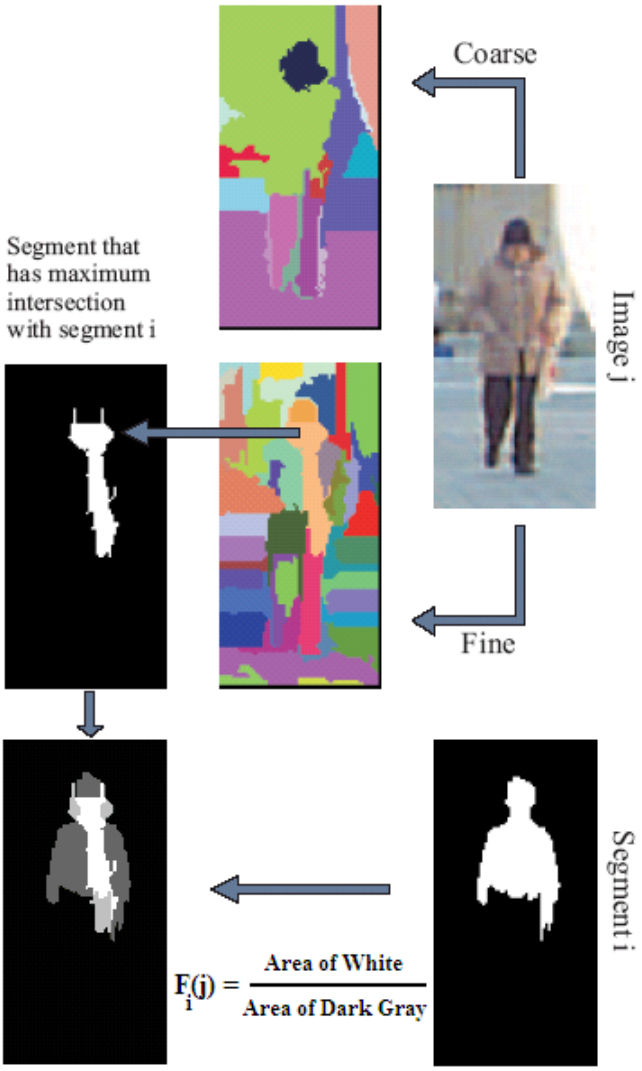


Fig. 4. Proposed method for calculating the feature vector $F_i(j)$. The left bottom image shows the result of intersecting segment *i* and the segment that has maximum intersection with that. The white area is the intersection part, the light gray area is the part of the intersecting segment that has no intersection with segment *i*, the dark gray area is the part of segment *i* that has no intersection.

E. Implementation

Due to computational complexity and memory constraints, instead of creating weak classifiers for all segment features in the training set and combining them to generate a strong classifier, the following process is implemented:

1) The training images are partitioned into 21 groups randomly,
2) The segment features and the weak classifiers for each group are created,
3) The best 1000 weak classifiers from each group is selected using the Adaboost algorithm,
4) The strong classifier is generated employing the Adaboost method for the second time on the best 21000 weak classifiers (21 groups × 1000 classifiers per group) of the whole training set.

Figure 5 illustrates the strong classifier obtained from the above process.

IV. Experimental Results

Note that the experimental results are NOT REAL RESULTS obtained from the proposed method because the results were not ready while writing this paper.

The performance of the method is evaluated on the INRIA dataset. INRIA dataset contains 1805 64×128 images of humans cropped from a varied set of personal photos and 1218 person free images in the training set.

The algorithm is implemented in C++ employing the OpenCV functions on a computer with an Intel Core 2, 6300 @ 1.86 MHz CPU and 2GB RAM.

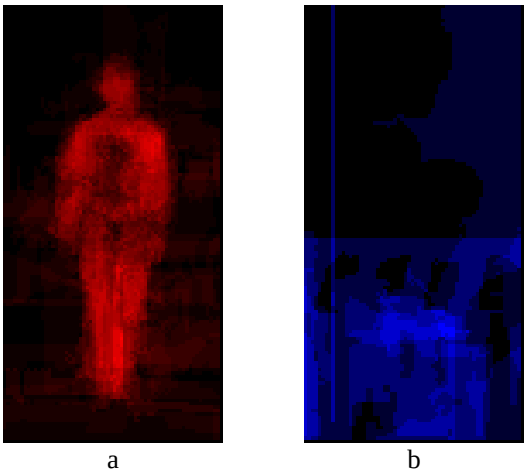
Figure 6 shows the comparison between the detection result of the proposed method and the detection results of two state-of-the-art methods in a Detection Error Tradeoff (DET) curve on a log-log scale (i.e., miss rate (8) versus False Positive Per Window (FPPW)).

$$Miss\ Rate = 1 - recall = \frac{False\ Negative}{True\ Positive + False\ Negative} \qquad (8)$$

Figure 6 shows that the proposed method outperforms the previous methods' miss rate by about 1% at FPPW of 10^{-4}.

In Figure 7, some falsely detected windows are depicted. Note that most of the false negatives are due to the unusual human pose (e.g., riding bicycle), semi occlusions, or the extremely light or dark environments.

Table 1 presents the processing time of the proposed method in different situations.



a b

Fig. 5. Strong classifier illustration, a) positive part, b) negative part.

V. Conclusion

In this paper, a novel feature-based pedestrian detection approach is presented. Unlike the current state-of-the-art methods [2],[3], which utilize some low-level features (image gradients) as the basic features, a mid-level feature, *segment feature*, is proposed that is capable of capturing more general image characteristics. To employ the combination of segment features, and also to make it computationally tractable, a two level boosting process is performed to create the final classifier.

Experimental results show that the proposed method outperforms the current state-of-the-art methods (1% less miss-rate in FPPW of 10^{-4}) in detection accuracy. However, its computational complexity demands some future works for real-time applications.

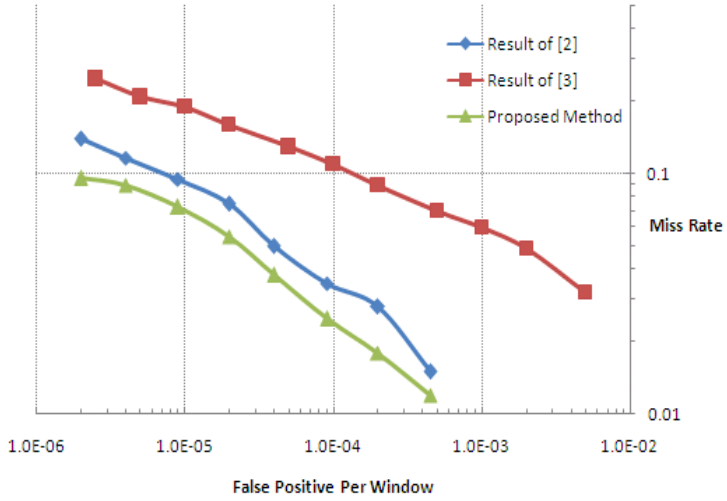


Fig. 6. Detection Error Tradeoff curve of the proposed method as well as two state-of-the-art methods.

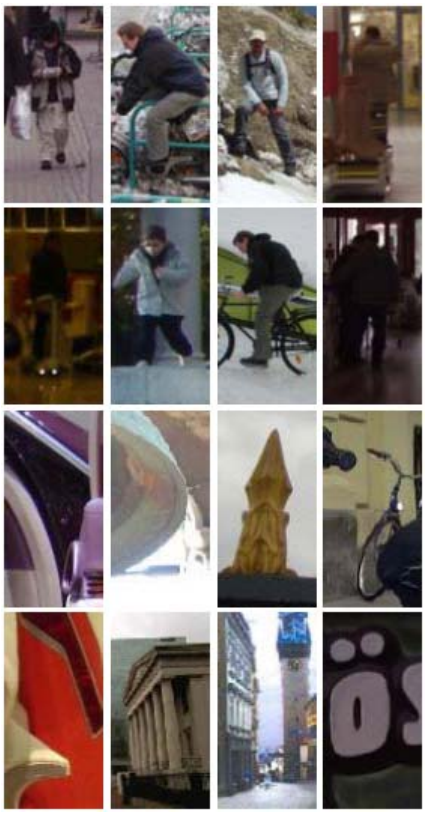


Fig. 7. Examples of falsely categorized windows. The two top rows are false negative examples and the two bottom rows are false positive examples.

TABLE I
TIME COMPLEXITY OF THE PROPOSED METHOD (SECONDS)

Image Size (pixel)	Number of utilized weak classifiers (T in Equation 7)		
	100	1000	10000
320×240	8.5	10	93
640×480	86	105	985

REFERENCES

[1] C. Papageorgiou, and T. Poggio, "A Trainable System for Object Detection," *Intl. Journal of Computer Vision*, vol. 38, no. 1, pp. 15-33, 2000.

[2] N. Dalal, B. Triggs, R. I. Alps, and F. Montbonnot, "Histogram of Oriented Gradients for Human Detection," *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, 2005.

[3] P. Sabzemeydani, G. Mori, "Detecting Pedestrians by Learning Shapelet Features," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[4] B. Leibe, E. Seeman, and B. Schiele, "Pedestrian Detection in Crowded Scene," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 878-885, 2005.

[5] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "pFinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, 1997.

[6] T. Zhao, and R. Nevatia, "Bayesian Human Segmentation in Crowded Situations," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 459-466, 2003.

[7] A. M. Elgammal, and L. S. Davis, "Probabilistic Framework for Segmenting People Under Occlusion," *Intl. Conf. on Computer Vision*, pp. 145-152, 2001.

[8] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *European Conference on Computational Learning Theory*, pp. 23–37, 1995.

[9] P. F. Felzenszwalb, and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *Intl. Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, 2004.

Mani Ranjbar was born in Semnan, Iran, in 1982. He received the B.Sc. degree in Computer Engineering from Sharif University of Technology, Tehran, Iran, in 2005. He joined the Image Processing Laboratory in the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, in 2005 and received his M.Sc. degree in that department in 2007. He is currently working toward his Ph.D. degree in the Vision and Media Laboratory, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada. His research interests include image processing, robotic vision and computer vision, focusing specifically upon object detection and recognition, scene recognition, segmentation, activity recognition, 3D object reconstruction, and tracking.

Multicasting JPEG2000 Images over MIMO Systems

Mahin Torki and Atousa HajShirMohammdi

***Abstract*—Increasing demand for high-speed and efficient multimedia transmission over wireless networks has driven tremendous research on enhancing the performance of multimedia communications over noisy channels. Multimedia applications increasingly require efficient transmission of still and moving images over wireless channels. In response to the rapidly increasing demand of the data-rate requirements, of particular importance is the so-called Multi-Input Multi-Output (MIMO) antenna. In this paper, we exploit spatial multiplexing, diversity, antenna selection, and Unequal Power Allocation (UPA) techniques simultaneously to transmit progressive JPEG2000 images over a MIMO downlink channel where communication occurs from a multi-antenna base station to many multi-antenna mobile terminals. We take advantage of the spatial diversity by utilizing a joint decoding in the JPEG2000 decoder. The aim is to reduce the average distortion among users. Simulation results show that our proposed algorithms provide significant image quality improvement when compared to schemes that do not consider antenna selection or UPA.**

I. Introduction

Real-time transmission of image and video content over wireless channels is becoming very common in cellular networks. The ultimate goal of future-generation wireless communications is to provide ubiquitous seamless connections between mobile terminals such as personal digital assistants and computer servers, so that users can enjoy high-quality multimedia services at anytime without wires. Fundamental physical challenges such as channel fading and interference, however, have put strains on the radio resources, which makes achieving reliable wireless communications difficult. Development of Multi-Input Multi-Outputs (MIMO) systems has been a great achievement toward overcoming this problem. These systems provide high speed links while maintaining good quality of service. MIMO systems have the following desirable capabilities: increased bitrate if used in spatial multiplexing systems, decreased bit error rate if used in diversity mode, and extended transmission range by utilizing beamforming methods [1]. On the other hand, JPEG2000 is the state-of-the-art still image compression standard that outperforms other image compression standards like JPEG [2].

In this paper, we consider multicasting of JPEG2000 images over MIMO systems. The JPEG2000 standard generates a progressive bitstream with different scalable progressions, excellent error resilience features, and region of interest processing. The property of JPEG2000 focused upon in this paper is its "quality scalability", which means as more elements of the bitstream are received, the quality will increase accordingly [2]. The aim of this paper is to present an antenna selection algorithm which transmits JPEG2000 images efficiently over a multiuser MIMO network, while making use of both spatial multiplexing and diversity benefits of MIMO systems. Due to the quality scalability of the JPEG2000 bitstreams, various parts of the encoded bitstream differ in importance, and it is crucial to assign the best channels in terms of Signal to Interference and Noise Ratio (SINR) to the most important parts in order to achieve the optimum distortion at the receiver.

Much research has been done considering optimized transmission of images and videos in wireless networks [3], [4], [5], [6], [7]. In [3] power efficient MPEG4-FGS video transmission over MIMO-OFDM systems is discussed. In [4] an unequal power allocation algorithm for JPEG transmission over MIMO systems is proposed. [8] proposes an energy efficient JPEG2000 image transmission system over point-to-point wireless sensor networks. In this work, we consider transmission of JPEG2000 images to multiple users over MIMO channels.

The challenge of antenna selection in a multicast MIMO system is because different users impose different transmitting antenna orders, and the best antenna selection strategy for one user may be the worst for another user. Our proposed algorithm assigns sub-streams to transmitting antennas to reduce the distortion of the received images for different users. To ensure that all users receive the important parts of the data through their best channels, the algorithm uses diversity and sends the important parts of the bitstream through multiple channels. Hence, the chance that users accurately receive the important parts of the bitstream, increases and also, we may use this extra information at the JPEG2000 decoder by jointly decoding the multiple copies of the same codestream. It is shown that applying antenna selection simultaneously for all the users, decreases the maximum distortion in the received images. Furthermore, wireless channels are subject to signal degradations such as noise, interference and fading and due to the nature of JPEG2000 coded bitstream, without adequate data protection, any transmission errors that occur in the coded image will be propagated to affect large image areas, causing visible and often objectionable, image quality deterioration. The JPEG2000 standard addresses the transmission error problem by including provisions for error resilience tools. Here we use the "error resilience" feature of JPEG2000 to improve the quality of received images.

The organization of this paper is as follows. In section II, we present our system model. In section III, the proposed method

Manuscript received March , 2008; revised April 3, 2008.

The authors are with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A1S6 (e-mail: mta33@sfu.ca,atousah@sfu.ca)

is explained, and section IV provides simulation results for investigating the performance of the proposed algorithm. Finally, section V discusses the results of the simulations.

II. System Description

The system is modeled as a two user MIMO, in which both users have the same requirements and have the same priority for the base station. All the transmitters and receivers have the same number of antennas, which we have chosen to be 4 in this paper. The general model of the system is shown in Fig. 1, with a description of different components given below.

A. Channel model

We assume that the channels between the base station and different users have the same statistics. Each of these channels is a 4 by 4 MIMO channel. The channel is assumed to be Rayleigh flat fading. The channel matrix entries are i.i.d. Gaussian complex random variables, with independent real and imaginary parts, each with zero mean and variance $1/2$. We assume that the channel is known at the transmitter and receiver. The noise is assumed to be AWGN with variance 1. We use 4-QAM for modulating the bitstream. The channel is slowly time varying and is assumed to be constant over every T symbol intervals.

B. Transmitter

The transmitter consists of two parts. The first part is the source encoder which converts the input images to compressed bitstreams. In this paper, the input images to the system are encoded using the JPEG2000 still image compression standard in the quality progressive mode. To encode the raw image, JPEG2000 first divides it into disjoint rectangular tiles. The subband/wavelet transform is applied to each tile-component to generate subbands, which are then divided into rectangular-shaped precincts, and further divided into square-shaped code-blocks. Each bitplane of a codeblock is then encoded by an arithmetic encoder in three codingpasses. This provides a progressive bitstream for each of the codeblocks. Coding passes are then interleaved to create the scalable JPEG2000 bitstream. We enable error resilient feature of the JPEG2000 encoder by using the RESTART/ERTERM option. We have used Kakadu as our JPEG2000 codec with 64×64 codeblocks and 128×128 precincts [9]. In all simulations, PPM markers were utilized, and the header information throughout the bitstream is separated and assumed to be transmitted error free. At the receiver, headers are re-inserted at their original location before JPEG2000 decoding.

The second part of the transmitter is the channel assignment unit. In this unit, after the removal of headers, the raw bitstream is divided into two equal length substreams, SS_1 and SS_2. Each substream is divides into non-overlapping blocks of lengths $2T$, where T is the number of symbols for which we assume the channel to be constant, and the number of bits per symbol for 4-QAM modulation is 2. Our proposed channel assignment algorithm then runs on each of these blocks independently. The total transmit power from all the antennas during each symbol period is kept constant at any given symbol interval. In effect, this unit determines the substreams that are transmitted through each antenna and the required power for transmitting them. It also determines how the final bitstream should be composed from the substreams that each user receives and sends this side information to the users. This part will be explained in more detail in section III-A.

C. Receiver

We have used a Minimum Mean Square Error (MMSE) receiver to eliminate the problem faced by MIMO receivers which is the presence of multi-stream interference. The MMSE is a linear receiver (i.e., it separates the transmitted substreams and then decodes each substream independently) [10].These substreams are then passed to our modified JPEG2000 decoder which will be discussed in section III-B.

III. Proposed Method

In this section, the channel assignment algorithm and our proposed modified JPEG2000 decoding scheme are discussed.

A. Channel Assignment Algorithm

The channel assignment algorithm, calculates the post processing Signal to Interference and Noise Ratios (SINR) for all the receiving antennas of each user every T symbols [11]. Because we have used MMSE receiver, each transmitted substream will be decoded from its corresponding receiver. The calculated SINRs determine the best transmitting antennas in terms of post processing SINR for each user. This constitutes the antenna selection order of each user. Based on the antenna selection order requested by each user, the algorithm assigns the antennas to substreams aiming at reducing the average distortion among all users. The algorithm is based upon the progressive nature of the JPEG2000 coded bitstream (i.e., the first substream needs more protection and should be transmitted through channels with the lower BERs). The challenge of antenna selection when transmitting images to multiple users will arise in cases that different users request different transmitting antenna orders, and the best antenna selection order for one user may result in great quality degradation in another user.

The algorithm assumes that the total transmit power at each symbol period is $4p$. According to the antenna selection orders requested by each user, the algorithm decides on one of the following scenarios: i) Choose two best antennas and send one copy of each substream with power $2p$ from each of the selected antennas. The other two antennas are not used for transmission. ii) Send two copies of each substream transmitting from all four antennas with power p. iii) Send one of the substreams with power $2p$ from one antenna and two copies of the other substream from two other antennas with power p. The fourth antenna is not used for transmission.

It should be noted that in cases where we send two copies of the same substream from two transmitting antennas, the

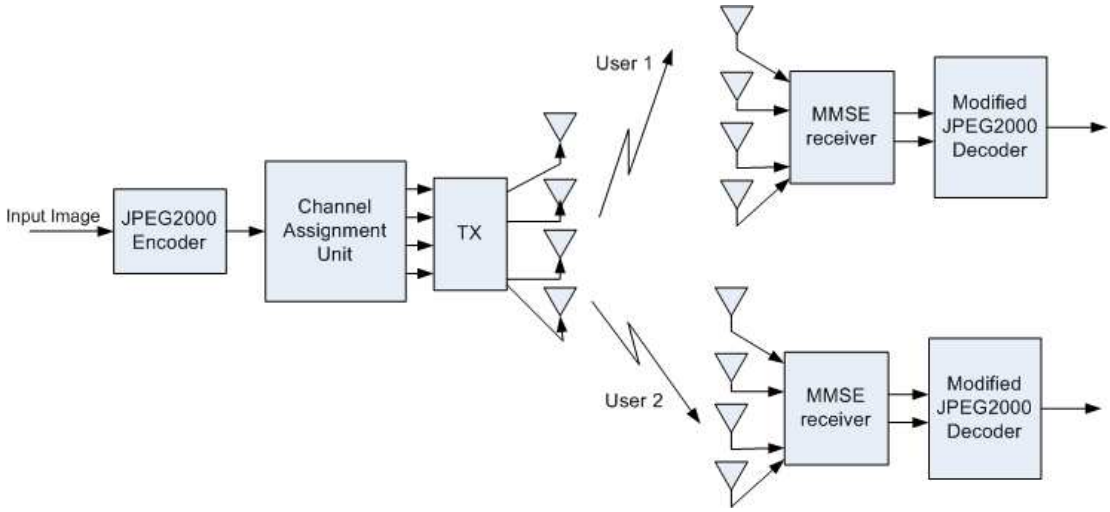


Fig. 1. System model for multicasting JPEG2000 images over MIMO channels

MMSE receiver will decode each copy separately and each user will receive two copies of that substream, obviously with different BERs. We call the copies transmitted over the channels with higher and lower SINR, the *main* and the *secondary* copy, respectively. The transmitter also lets each user know which substream has been transmitted from each antenna.

The channel assignment algorithm is summarized in **Algorithm 1**, where SS_{k1} and SS_{k2} refer to the two copies of the k^{th} bitstream, $k = 1, 2$, and A_{ij} refers to the j^{th} best antenna of user i, $i = 1, 2$, $j = 1, \ldots, 4$.

Algorithm 1 Channel Assignment Algorithm

for Every T symbols **do**
Calculate the post processing SINR for all the receiving antennas of each user and sort the antennas based on their corresponding SINRs .

if The two best antennas of each user match **then**
Power $(SS_1) = 2p$
Power $(SS_2) = 2p$
Transmit SS_1 from A_{11}
Transmit SS_2 from A_{12}
Do not transmit from remaining antennas
end if

if Only the best antennas of each user match **then**
Power $(SS_1) = 2p$
Power $(SS_{21}) = p$
Power $(SS_{22}) = p$
Transmit SS_1 from A_{11}
Transmit SS_{21} from A_{12}
Transmit SS_{22} from A_{22}
Do not transmit from the remaining antenna
end if

if The best antennas of the two users are different **then**
Power $(SS_{11}) = p$
Power $(SS_{12}) = p$
Power $(SS_{21}) = p$
Power $(SS_{22}) = p$
$CH1_1 \Leftarrow SS_{11}$
$CH2_1 \Leftarrow SS_{12}$
Send SS_{21} and SS_{22} from the remaining antennas
end if
end for

B. Modified JPEG2000 decoder

In the original JPEG2000 decoder (in the ERTERM/RESTART mode) if an error is detected in a codingpass, all the remaining coding passes from the current codeblock are discarded [9]. The modified JPEG2000 decoder, however, takes advantage of the diversity in the received data during the time intervals that more than one copy of codestream is available. This is done by jointly decoding the two different received copies of the original data. At first, the decoder uses the *main* codestream to decode the image. Once the decoder detects an error in a coding pass, by employing **Algorithm 2**, the decoder tries to correct the erroneous byte(s) by using the information contained in the *secondary* codestream. If the decoder is successful in correcting the damaged codingpass, it restarts its operation from the beginning of the erroneous codingpass. Otherwise, it operates normally. The modified JPEG2000 decoder is summarized in **Algorithm 2**.

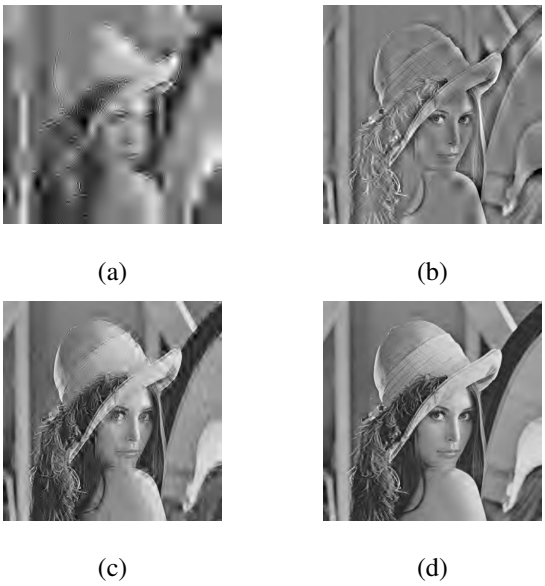
IV. Simulation Results

In this section, we provide experimental results to investigate the performance of our proposed algorithms. Images

Algorithm 2 Modified JPEG2000 Decoder

```
for All the codingpasses in the codestream do
    Decode the codingpass.
    if Decoder declares occurrence of an error in the codingpass then
        Find all the mismatches between the main and the secondary codingpasses

        for each mismatch do
            Swap the corresponding values of the two codestreams
            Reinitialize the decoding from the start of the codingpass
            if decoding is successful then
                Break;
            else
                swap the values of the two codestreams again and repeat the procedure with the next mismatch
            end if
        end for
    end if
end for
```



(a) (b) (c) (d)

Fig. 2. Lena image results for different schemes at SNR=15 dB (a) Received image by User 2 with antenna selection based on User1, 1.0 bpp, (b) Received image by User 1 with antenna selection based on User1, 1.0 bpp (c) EPA-JAAMD, 0.5 bpp, (d) UPA-JAAMD, 0.5 bpp

are encoded with 4 different quality layers with the quality progression option enabled in the JPEG2000 encoder. Also, to take advantage of the error resilient feature of JPEG2000, we enable the ERTERM/RESTART mode at the encoder. Consequently, the decoding should be done in the error resilient mode, as well. We assume that the channel is constant for 250 symbols to ensure that the slowly time varying condition is satisfied [4]. The results shown here are for the 512×512 Lena image with a source coding rate of 1.0 bpp.

We compare the performance of our proposed algorithms with two more cases. The results are shown in Fig. 3.

In the first case, the image is encoded with 1.0 bpp and is divided into 4 equal length substreams. We transmit the four substreams according to the antenna selection order requested by user 1. In effect, we are giving complete priority to the needs of user 1 and we are applying no antenna selection for user 2. We found the PSNR curves for User1 and User2, labeled with "User1 based-U1" and "User1 based-U2", respectively. We also calculated the average PSNR for the two users, labeled with"Ave U1 and U2" in the figure.

In the second case, the input image is encoded using only 0.5 bpp and is divided into two equal length substreams which are sent according to our proposed algorithm. This results in the same transmission time and power consumption as the previous two cases for a fair comparison. We call our proposed method, "Joint Antenna Assignment and Modified JPEG2000 Decoder" (JAAMD). Since we use different powers for transmitting the substreams, the corresponding performance curve is labeled as "UPA-JAAMD" in the figure, to signify the Unequal Power Allocation employed here.

In the third case, we use our proposed method at the same rate (0.5 bpp) with the difference that we always transmit from all four antennas with equal power. We call this method "EPA-JAAMD", where EPA stands for Equal Power Allocation.

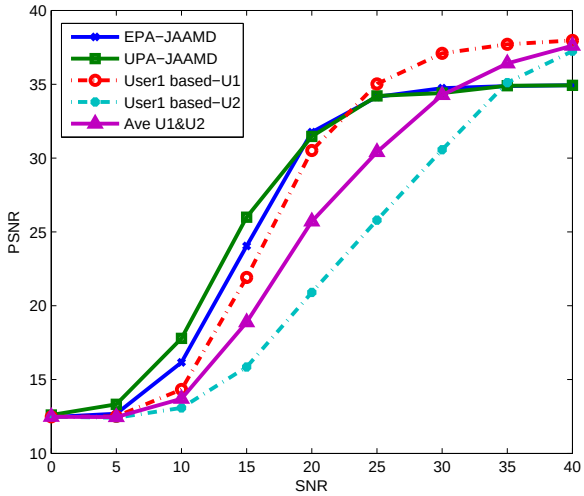


Fig. 3. PSNR curves for different schemes

To show the visual quality of our proposed method, results for Lena image transmitted with different algorithms at 15 dB are shown in Fig. 2. Noticeable visual quality enhancement is achieved through our proposed algorithms. As can be seen from Fig. 3, our algorithm performs significantly better in terms of PSNR in the SNR range of $0 - 30$ dB compared to the case we send the 1.0 bpp bitstream without antenna selection or based on the required order of one of the users. Simulation results show that for low SNR conditions, it is more beneficial to send the more important parts of the coded image with great protection, as these parts construct the main part of the decoded image PSNR.

V. Conclusion

In this paper, we presented an algorithm for multicasting of JPEG2000 images over MIMO systems. According to different antenna selection orders requested by different users, the channel assignment unit decides to transmit the substreams with different power levels from the antennas. The decoder is modified such that, when two copies of a substream are available, they are used jointly to decode the image, using the error resilient feature of JPEG2000 decoder. The proposed method, takes advantage of spatial multiplexing and diversity in MIMO systems, while applying UPA. Simulation results show that our algorithm provides a significant gain over other presented schemes. Future research direction is to extend this algorithm to more than two users.

References

[1] D. Gesbert, M. Shafi, D. Shiu, P.J. Smith, and A. Naguib, "From theory to practice: an overview of mimo space-time coded wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3, pp. 281–302, 2003.

[2] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 36–58, 2001.

[3] Z. Ji, Q. Zhang, W. Zhu, Z. Guo, and J. Lu, "Power-efficient mpeg-4 fgs video transmission over mimo-ofdm systems," *Communications, 2003. ICC'03. IEEE International Conference on*, vol. 5, 2003.

[4] M.F. Sabir, R.W. Heath Jr, and A.C. Bovik, "Unequal power allocation for jpeg transmission over mimo systems," .

[5] S. Appadwedula, M. Goel, DL Jones, K. Ramchandran, and NR Shanbhag, "Efficient wireless image transmission under a total power constraint," *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 573–578, 1998.

[6] S. Appadwedula, M. Goel, N.R. Shanbhag, D.L. Jones, and K. Ramchandran, "Total system energy minimization for wireless image transmission," *The Journal of VLSI Signal Processing*, vol. 27, no. 1, pp. 99–117, 2001.

[7] D.G. Sachs, R. Anand, and K. Ramchandran, "Wireless image transmission using multiple-description based concatenated codes," *Proc. SPIE Image Video Processing*, vol. 3974, pp. 300–311, 2000.

[8] W. Yu, Z. Sahinoglu, and A. Vetro, "Energy efficient jpeg 2000 image transmission over wireless sensor networks," *Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE*, vol. 5.

[9] D.S. Taubman and M.W. Marcellin, *Jpeg2000: Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2002.

[10] A. Paulraj, D. Gore, and R. Nabar, *Introduction to space-time wireless communications*, Cambridge University Press, 2005.

[11] R.W. Heath Jr, S. Sandhu, and A. Paulraj, "Antenna selection for spatial multiplexing systems with linearreceivers," *Communications Letters, IEEE*, vol. 5, no. 4, pp. 142–144, 2001.

Mahin Torki received her B.Sc. degree from Department of Electronics, Faculty of Computer and Electrical Engineering, Isfahan University of Technology (IUT), Iran, in 2004. During the summers of 2002-2004, she was a summer researcher at Artificial Intelligence Lab (AIL) and Operating System Lab (OSL) at IUT. From 2004-2006, she worked as a circuit designer for Sub-marine R&D Center. Since May 2007, she has been a master student in Electrical Engineering, Department of Engineering Science, Simon Fraser University (SFU), Burnaby, BC, Canada. Her research interests are in wireless communications, MIMO systems, joint source-channel coding, and multimedia transmission over wireless networks.

Data Throughput Analysis for Rapid Dynamic Spreading Gain Adaptation in CDMA Networks with Imperfect Channel Information

Song Xue

Abstract—**This paper analyzes the throughput of symbol-by-symbol duration adaptation. Satisfying a certain symbol error probability (SEP) constraint, the channel gain thresholds are optimized to achieve maximum throughput. The results show that, with perfect channel information, symbol-by-symbol (SBS) duration adaptation achieves a significant throughput gain over frame-by-frame (FBF) adaptation schemes for fast fading channels – in the orders of magnitude. And also higher throughput gain can be achieved by applying more levels of duration to adapt wireless channels. On the other hand, if only imperfect channel information is available, the throughput gain is still promising, although the imperfect channel information has important impact on the performance of adaptive transmission.**

Index Terms—**symbol rate adaptation, throughput**

I. INTRODUCTION

THE increasing demand for voice, data and multimedia service propels the need for higher capacity and data rates for wireless communication. However, the radio spectrum is a limited resource, which is unable to provide an arbitrary amount of wireless channels or bandwidth. Thus improving the radio spectral efficiency is one of the most important methods to meet the growing need for communication capacity. The spectral efficiency measures the amount of information bits transferred over a given bandwidth, and its unit is $bit/s/Hz$. If the channel bandwidth is fixed, increasing the transmission rate is equivalent to improving the spectral efficiency.

To increase the symbol (bit) rates over wireless channels, many adaptation techniques have been proposed to deal with channel fading. Adaptive constellation size assigns a varied number of information bits to one transmitted symbol, so that higher information throughput can be achieved at the same symbol rate. Adaptive power control, which sends signals with power as low as possible in order to reduce multi-user interference and error probability, also leads to higher system throughput. In CDMA networks, the symbol duration can be adjusted by applying spreading codes with different code lengths. This method is symbol duration adaptation which can improve the spectral efficiency for CDMA systems.

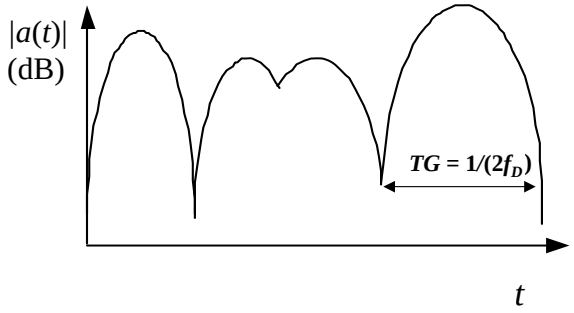


Fig. 1 Time-varying fading channel gain

Wireless channel fading is a phenomenon where instantaneous channel gains fluctuations are caused by multi-path propagation. Due to the presence of reflectors in wireless communication environments, the signals arrive at the receiver with multiple copies, and each suffers different time delays, phase shifts, and amplitude attenuations. The superposition of these copies may cause either constructive or deconstructive interference to the received signal power. The changing rate of amplitude attenuation and phase shift is related to the velocity of mobiles and reflectors around the mobiles. A moving mobile experiences a frequency shift away from its carrier frequency, which is known as Doppler shift. Doppler shift is measured as:

$$\Delta f = f_c \frac{v}{c}, \tag{1}$$

where f_c is the carrier frequency, v is the velocity of the mobile, and c is the propagation speed of the carrier, which is the speed of electromagnetic waves $(3\times10^8 m/s)$. As shown in Fig. 1[1], the fluctuation of wireless channel gain is more or less periodical, where period T_G is approximately

$$T_G = \frac{1}{2 f_D}, \tag{2}$$

where f_D is the maximum Doppler shift.

To maintain the quality of communication, signals must be transmitted with sufficient power and a under certain rate to obtain adequate signal to noise ratios (SNR) on received signals which have suffered fading. Satisfying a certain error probability requirement, the symbol duration (rate)

adaptation is to track the channel gain fluctuation and send signals with a rate as high as possible so that more information can be transmitted.

Conventionally, duration adaptation adjusts the symbol duration (rate) only at the beginning of each frame, which is identified as frame-by-frame (FBF) duration adaptation in this paper. When the mobile speed is relatively high, the channel fluctuation period, T_G, is close to, or even less than the frame duration. For example, suppose the frame duration is 10ms, if the mobile is moving at a speed of 36 km/h, and the carrier frequency is 1.5GHz, by equation (2), the $T_G = 10ms$ as well. In this case, the signals suffer one or several periods of channel fluctuation. Thus, the symbol duration adjustment is not able to track the channel gain varying but can only take the average channel gain into account. Therefore, for fast time-varying channels, the FBF duration adaptation is not beneficial to combating fading[2].

Compared to FBF duration adaptation, a symbol-by-symbol (SBS) adaptation scheme[2] is much more efficient for dealing with fast time-varying channels. SBS duration adaptation changes the symbol rate for each symbol instead of for each frame, providing the capability to adapt the channel fluctuation precisely and yield much higher average transmission rates and spectral efficiencies.

In order to perform symbol duration adaptation, or more generally, to perform any kind of adaptive transmission, the channel information, often called as channel state information (CSI), must be known by both the transmitter side and the receiver side. The channel gain, which is the CSI needed for symbol duration adaptation, is estimated by the receiver and sent back to the transmitter through a feedback channel. The process of estimation and feedback introduce time delays , and the time delays make the CSI used by the transmitter outdated. So the transmitter must predict the current channel gain from the past estimation. The imperfection of an outdated CSI has a significant impact on adaptation performance[3].

This paper analyzes the throughput gain of the SBS duration adaptation over the FBF method. In section II, the SBS scheme is introduced, and its throughput gain is analyzed in section III. An essential analysis result of the outdated CSI impact is shown in section IV.

II. SYMBOL BY SYMBOL DURATION ADAPTATION

In Direct Sequence CDMA (DS-CDMA) networks, the baseband information signals are multiplied by the wideband code signal to spread the signal spectrum[4]. The spread-spectrum signals are then modulated and sent through the wireless channel. On the receiver side, the received signals need to be demodulated and also de-spread to recover the baseband information signal. This process is named spread-spectrum because, multiplied by the direct sequence spreading code, the outcome signals occupy a much larger bandwidth than the original baseband information signals.

CDMA chip duration, T_C, is defined as the bit interval of the wideband direct sequence codes, while the bit interval of the information codes is denoted as T_b. The ratio of the two is called the spreading gain G_P:

$$G_P = \frac{T_C}{T_b}, \tag{3}$$

The duration (rate) adaptation is the method of using a set of spreading codes with different code lengths to adjust the symbol transmission rate. With fixed chip duration T_C, multiplied by different length spreading code, the resulting spread-spectrum symbols have different symbol duration (rate).

The symbol by symbol duration adaptation employs orthogonal varied Spread Factor (OVSF) code sequence sets, in which the codes have different lengths and are shift orthogonal to each other. Suppose $S = \{C_1, C_2, C_3, \cdots, C_N\}$ is an OVSF code sequence set, then S has the following properties:

1) Each term of the code sequence takes a value of 1 or -1.
2) The code sequences have the lengths of a power of 2 (i.e., if the length of code sequence, C_i ,is denoted as $|C_i|$, then $|C_i| = 2^{i-1}|C_1|$).
3) All sequences are shift orthogonal to each other.

Shift orthogonality is defined as that code sequence, U ,of length, n, and code sequence V of length nL are shift orthogonal with the length n if and only if

$$\sum_{k=1}^{n} U[k]V[nl+k] = 0, \tag{4}$$

for each $l = 0,1,2,\cdots,L-1$ [2].

The advantage of applying an OVSF code sequence set in the symbol duration adaptation system is that the receivers can determine which code sequence should be used to de-spread each received symbol without any message from the transmitters. By correlating the received symbols with each sequence of the OVSF code set, the receivers determine that the sequence with non-zero outcome is the one which the transmitter used for spreading the current symbol. This advantage is crucial in symbol by symbol duration adaptation because extra messages are avoided. Since the code sequences used for each information symbols are different, if the messages are needed by receivers to choose the right sequences, an extra message must be sent by the transmitters for each symbol. Obviously, those additional

messages occupy a large proportion of channel transmission time and decrease the channel information throughput and spectral efficiency as well.

III. THE THROUGHPUT OF SYMBOL BY SYMBOL DURATION ADAPTATION

As introduced in section II, the symbol by symbol duration adaptation is performed by applying a set of OVSF code sequence sets. By using the OVSF sequence set:

$$S=\{C_1,C_2,C_3,\cdots,C_N\},$$

the transmitters sends symbols with a set of discrete rates:

$$\{r_1,r_2,\cdots,r_N\},$$

where $r_i=\dfrac{1}{|C_i|T_C}$ and also $r_i=2r_{i+1}$.

Hence, the discrete rate set actually is:

$$\left\{r_1,\frac{r_1}{2},\frac{r_1}{4},\cdots,\frac{r_1}{2^{N-1}}\right\},$$

The transmission rate is adjusted in accordance with the wireless channel gain, denoted as α. In this section, the instantaneous channel gain is assumed to be known for both the transmitter and receiver sides, (i.e., perfect CSI is known).

As shown in Fig 2, the range of fading power gain, $|\alpha|^2$, is divided to N regions with N-1 thresholds $\{\zeta_1,\zeta_2,\cdots,\zeta_{N-1}\}$. The symbols are sent with the rate, r_i,when the power gain satisfies $\zeta_{i-1}\le|\alpha|^2<\zeta_i$. Then the average rate is:

$$\overline{r_s}=\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} f_{|\alpha|^2}(x)dx, \tag{5}$$

where $f_{|\alpha|^2}(x)$ is probability distribution of channel power gain, $|\alpha|^2$, and $\zeta_0=0$, $\zeta_N=\infty$ .If the symbol error probability of the information bits is function $h(\gamma)$, where γ is the signal-to-noise ratio (SNR) of the information bits, the average error probability with the symbol duration adaptation is:

$$\overline{SEP}=\frac{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} h(\frac{Px}{N_0r_i})\,f_{|\alpha|^2}(x)dx}{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} f_{|\alpha|^2}(x)dx}, \tag{6}$$

where P is the transmission power, and N_0 is the power spectral density of the additive white Gaussian noise. Satisfying an error probability constraint, $\overline{SEP}\le\varepsilon$, an optimal threshold set is required to achieve a maximal throughput(i.e., maximal average rate $\overline{r_s}$). Thus, the optimization is formulated as follows:

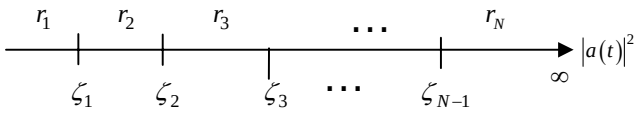


Fig.2 Channel Power Thresholds for Symbol Rate Adaptation

$$\max_{\zeta_1,\zeta_{2,\cdots},\zeta_{N-1}} \sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} f_{|\alpha|^2}(x)dx,$$

subject to

$$\frac{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} h(\frac{Px}{N_0r_i})\,f_{|\alpha|^2}(x)dx}{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} f_{|\alpha|^2}(x)dx}\le\varepsilon, \tag{7}$$

To compare the throughputs between SBS and FBF duration adaptation, throughput gain is defined as the ratio of their average rates:

$$R=\frac{\overline{r_S}}{\overline{r_F}}, \tag{8}$$

In this paper, the channel is modeled as a Rayleigh fading channel (i.e., the normalized fading power gain $|\alpha|^2$ has an exponential distribution):

$$f_{|\alpha|^2}(x)=\exp(-x),x>0 \tag{9}$$

To show the performance of SBS duration adaptation, let BFSK (binary frequency shifting key) modulation be the example. For BFSK, the error probability function is:

$$h(\gamma)=\frac{1}{2}\exp\left(-\frac{\gamma}{2}\right), \tag{10}$$

where γ is the signal-to-noise ratio (SNR) of the information bits. Substitute equation (9), (10) into (7), and the maximum average symbol rate is:

$$\overline{r_s}=\max_{\zeta_1,\zeta_{2,\cdots},\zeta_{N-1}} \sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} \exp(-x)dx,$$

subject to

$$\frac{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} \frac{1}{2}\exp(-\frac{Px}{N_0r_i})\exp(-x)dx}{\sum_{i=1}^{N} r_i\int_{\zeta_{i-1}}^{\zeta_i} \exp(-x)dx}\le\varepsilon, \tag{11}$$

The average symbol rate of FBF duration adaptation for BFSK has been derived by[5]:

$$\overline{r_F}=\frac{P}{N_0(\frac{1}{\varepsilon}-2)}, \tag{12}$$

Combined with equation (11) and (12), the throughput gain (8) can be numerically computed and the result is shown in Fig.3. The figure indicates that the throughput gain

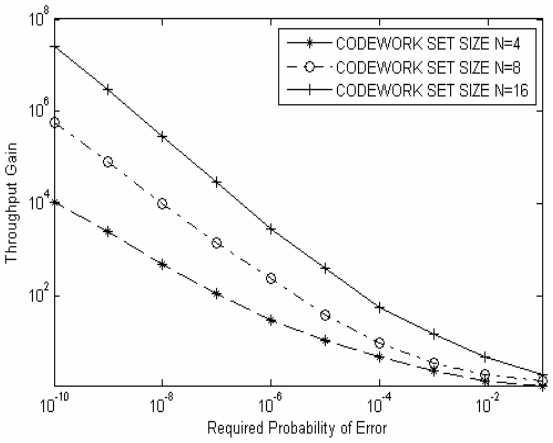


Fig. 3 Throughput Gain of BFSK

is achieved by SBS and the gain increases in orders of magnitude as the error constrain, ε, decreases.

The number of code sequences applied in adaptation is also interesting and the simulation results are also shown in Fig. 3. The results show that the more sequences used in adaptation, the higher throughput gain achieved. This result agrees with intuition and the reason for it is that the system can adapt the channel more precisely by applying more levels of transmission rates, while the cost must be paid for the increasing system complexity.

IV IMPACT OF IMPERFECT CHANNEL INFORMATION ONTHROUGHPUT GAIN

Reference [3]has presented the quantitative analysis of the throughput gain with the outdated channel estimation, and some essential results are given here.

The wireless channel is modeled as an isotropic scattering channel[3], in which channel gain, $a(t)$,is assumed to be an ergodic zero-mean complex random process with autocorrelation:

$$E\left[a(t+\tau)a(t)^*\right] = E\left[|a(t)|^2\right] J_0(2\pi f_D \tau),\ \forall t,\tau, \qquad (13)$$

where J_0 is zero-order Bessel function of the first kind. The model implies that $a(t)$ is a circular process as well[6].

$$E\left[a(t+\tau)a(t)\right] = 0,\ \forall t,\tau, \qquad (14)$$

As the transmission power is denoted as P , at time, t, the received signal power is $P|a(t)|^2$ and the signal-to-noise ratio (SNR) is

$$SNR = \frac{P|a(t)|^2}{N_0 r\left(|a(t)|^2\right)}, \qquad (15)$$

where $r\left(|a(t)|^2\right)$ is the symbol rate when the power gain is $|a(t)|^2$, as $r(\bullet)$ denotes the symbol rate adaptation policy. To simplify the analysis, a rudimentary linear policy is employed:

$$r\left(|a(t)|^2\right) = qP|a(t)|^2, \qquad (16)$$

where q is a constant coefficient that needs to be determined.

As described in introduction, the transmitters can only obtain the delayed channel estimation and must predict the current channel power gain from the outdated channel estimation. Let $|\hat{a}(t)|^2$ denote the estimation of channel power gain at time, t, $|a(t)|^2$. Suppose the system has time delay, τ ,between the estimation at the receiver side and rate adjustment at the transmitter side, the channel power gain at time, $t+\tau$,is denoted as $|a(t+\tau)|^2$ and its prediction based on $|\hat{a}(t)|^2$ is $|\hat{a}(t+\tau)|^2$ .Hence, at time, $t+\tau$, the symbol rate is:

$$r\left(|\hat{a}(t+\tau)|^2\right) = qP|\hat{a}(t+\tau)|^2, \qquad (17)$$

and the SNR is

$$\underset{t+\tau}{SNR} = \frac{P|a(t+\tau)|^2}{N_0 r\left(|\hat{a}(t+\tau)|^2\right)}, \qquad (18)$$

The average symbol rate is

$$\overline{r_s} = E\left[r\left(|a(t)|^2\right)\right], \qquad (19)$$

and average error probability is

$$\overline{SEP} = \frac{E\left[r\left(|\hat{a}(t+\tau)|^2\right) h\left(\frac{P|a(t+\tau)|^2}{N_0 r\left(|\hat{a}(t+\tau)|^2\right)}\right)\right]}{E\left[r\left(|a(t)|^2\right)\right]}, \qquad (20)$$

Where $h(\bullet)$ denotes the symbol error probability of the information bits.

The optimal q is determined by maximizing the average rate, $\overline{r_s}$,while the average error probability, $\overline{SEP} \le \varepsilon$, where ε is the fidelity constraint of error probability. In order to solve the optimization problem, the relation between $|a(t+\tau)|^2$ and $|\hat{a}(t+\tau)|^2$ must be derived. As $a(t)$ and its estimation error, $a_v(t)$,are modeled as statistically independent and Gaussian processes, and $\hat{a}(t)$ and $a(t+\tau)$ are assumed to be jointly Gaussian, $|a(t+\tau)|^2$ and $|\hat{a}(t+\tau)|^2$ are also joint Gaussian, form a zero-mean complex Gaussian random vector[3]. The average rate is

$$\overline{r_s} = qP\Upsilon(\tau)E\left(|a(t)|^2\right), \qquad (21)$$

where

$$\Upsilon(\tau) = \frac{E\left(|a(t)|^2\right)}{E\left(|a(t)|^2\right) + E\left(|a_v(t)|^2\right)}\left(J_0(2\pi f_D \tau)\right)^2, \qquad (22)$$

Let BFSK be the example, so

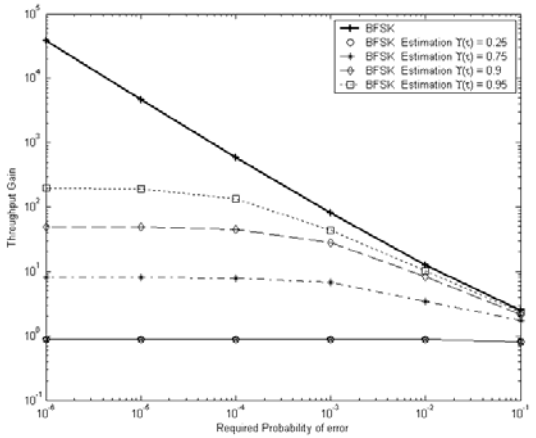


Fig. 4 Throughput Gain with Imperfect CSI

$$h(\gamma) = \frac{1}{2}\exp\left(-\frac{\gamma}{2}\right)$$

q is optimized by solving

$$\max_q \overline{r_s} = \max_q qP\Upsilon(\tau)E\left(|a(t)|^2\right),$$

subject to

$$\frac{E\left[|\hat{a}(t+\tau)|^2 \frac{1}{2}\exp\left(-\frac{1}{2}\frac{|a(t+\tau)|^2}{N_0 q|\hat{a}(t+\tau)|^2}\right)\right]}{E\left[|\hat{a}(t+\tau)|^2\right]} \le \varepsilon, \qquad (23)$$

Then the throughput again for BFSK is

$$R = \frac{\overline{r_s}}{\overline{r_F}} = 2N_0 q\Upsilon(\tau)\left(\frac{1}{2\varepsilon}-1\right), \qquad (24)$$

The numerical computation results are shown in Fig.4. In the figure, the top solid line is the ideal gain when perfect CSI is assumed to be known. Other curves show that the throughput gains are degraded while $\Upsilon(\tau)$ decreases. Shown in the equations (22) and (13), $\Upsilon(\tau)$ is linearly related to the squared autocorrelation function of channel gain, $a(t)$, and it well demonstrates the imperfection of the CSI from the time delay, τ. However, with imperfect CSI, the SBS duration adaptation still shows promising throughput gain over the FBF scheme.

V. CONCLUSION AND FUTURE WORK

In this paper, the throughput gain of symbol-by-symbol duration adaptation over frame-by-frame duration adaptation is analyzed. The results show, when ideal channel information is available, that SBS adaptation achieves significant throughput gain and improves the channel spectral efficiency remarkably. Even with outdated channel information, employing a rudimentary adaptation policy and basic prediction scheme, SBS adaptation still has promising performance. Future research will develop more precise estimation-feedback-prediction schemes and adaptation policies to deal with the imperfect channel information.

REFERENCES

[1] W. C. Jakes, *Microwave Mobile Communication*, New York: IEEE Press, 1993.

[2] L. Tsaur and D. C. Lee, "Closed-loop architecture and protocols for rapid dynamic spreading gain adaptation in CDMA networks," *IEEE/ACM Transactions on Networking*, vol. 14, no. 4, August 2006, pp. 821–834.

[3] D. C. Lee, M. Naeem, and S. Xue, "On Error-Constrained Data Throughput of Rapid Symbol Duration Adaptation with Imperfect Channel Information, " to be published.

[4] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communication* Reading, MA: Addision-Wesley, 1995.

[5] D. C. Lee and L. Tsaur, "Gains Achieved by Symbol-by-Symbol Rate Adaptation on Error-Constrained Data Throughput over Fading Channels," *Journal of Communication and Networks*, vol. 9, no. 3, September 2007, pp. 240–245.

[6] J. G. Proakis, *Digital Communications*, 4th Ed. Boston: McGraw-Hill, 2000.

Song Xue received his BS degree in applied science from the University of Electronic Science and Technology of China in 1990. He worked for the Institute of Telecommunication of Sichuan Province, China as a network engineer from 1990 to 1996. Then he joined SSK System Inc., and worked as a network engineer until 2002. He is currently a Master candidate in School of Engineering , Simon Fraser University, Canada.