OPTIMIZED TRANSMISSION OF H.26L/JVT CODED VIDEO OVER PACKET-LOSSY NETWORKS

Thomas Stockhammer
Institute for Communications Engineering
Munich University of Technology
80290 Munich, Germany

Thomas Wiegand
Image Processing Department
Heinrich-Hertz- Institute,
10587 Berlin, Germany

Stephan Wenger
Communication and Operating Systems
Technical University Berlin
10587 Berlin, Germany

ABSTRACT

Transmission of hybrid coded video including motion compensation and spatial prediction over error-prone channels results in the well-known problem of spatio-temporal error propagation at the decoder. A widely accepted standard-compliant technique to enhance the quality of the decoded video significantly is the more frequent introduction of intra-coded macroblocks. However, intra-coded information generally requires more bit rate. Therefore, a careful selection of intra-updates in terms of rate and distortion is necessary. A flexible and robust rate-distortion optimization technique is presented to select coding mode and reference frame for each macroblock. The channel statistics are included in the optimization process. We derive a method to obtain an estimate of the decoder pixel distortion at the encoder. The presented techniques are verified within the new H.26L/JVT video coding standard based on common test conditions.

1 INTRODUCTION

The transmission of video over packet lossy network is more important than ever before. Motivated by the increasing traffic on packet switched networks like ATM or the Internet, there is tremendous interest for robust image and video transmission for lossy packet networks. Though the theoretical framework for packet lossy transmission is well-known under the acronym multiple description coding [1], the application to video transmission still seems to be an open problem. The highly complex temporal and spatial prediction mechanisms included in modern video codecs like JVT/H.26L [2][1] coding result in catastrophic error propagation in case of packet losses. Transmission errors could be reduced by appropriate channel coding techniques. For channels without memory, such as the AWGN channel, channel-coding techniques provide very significant reductions of transmission errors at a comparably moderate bit-rate overhead. For the mobile fading channel [3] and the Internet [4], however, the effective use of forward error correction and re-transmission is limited when assuming a small end-to-end delay. Here, the use of error resilience techniques in the source codec becomes important.

In standardized video decoders like H.26L/JVT coding quick recovery can only be achieved when image regions are encoded in Intra mode, i.e., without reference to a previously coded frame. The Intra mode, however, is not selected very frequently during normal encoding and completely Intra coded frames are not usually inserted in real-time encoded video as is done for storage or broadcast applications. Instead, only single macroblocks are encoded in Intra mode for regions that cannot be predicted efficiently. Conservative approaches transmit a number of Intra coded macroblocks anticipating transmission errors. In this situation, the selection of Intra coded macroblocks can be done either randomly or preferably in a certain update pattern. For example, Zhu [2] has investigated update patterns of different shape, such as 9 randomly distributed macroblocks, 1x9, or 3x3 groups of macroblocks. Although the shape of different patterns slightly influences the performance, the selection of the correct Intra percentage has a significantly higher influence. In [6] and [7], it is shown that it is advantageous to consider the image content when deciding on the frequency of Intra coding. For example, image regions that cannot be concealed very well should be refreshed more often, whereas no Intra coding is necessary for completely static background.

More recent work considers the use of Lagrangian macroblock mode decision [8][9] when assigning Intra macroblocks with significant improvements in rate distortion performance. In [10][11][12], random reconstruction results at the decoder side are considered which depend on the statistics of the transmission errors that cause a concealment and the motion compensation that determines the inter-frame error propagation. The reconstruction quality at the decoder, i.e., the average decoding distortion, is determined by the source coding distortion, which quantifies the error between the original signal and the reconstructed signal at the encoder, and the divergence between encoder and decoder.

In this work we will present on a similar method to obtain a rate-distortion optimized H.26L encoder for packet lossy networks. However, a different complex but robust approach to estimate the decoder distortion is introduced. The good performance of this new algorithm and the suitability of H.26L/JVT for packet lossy networks are verified by experimental results.

2 PACKET LOSS-OPTIMIZED ENCODER

2.1 Problem Formulation

The investigated video transmission system is shown in Figure 1. H.26L/JVT video encoding is based on a sequential encoding of frames denoted with the index $n, n = 1,\ldots,N$ with N the total number of frames to be encoded. In most existing video coding standards including H.26L, within each frame video encoding is typically based on sequential encoding of macroblocks denoted by index $m, m = 1,\ldots,M$ where M is total number of macroblocks in one frame and depends on the spatial resolution of the video sequence. Macroblocks are generally quadratic with size $\sqrt{I} \times \sqrt{I}$ pixel, i.e. one macroblock contains

[1] All referenced standard documents can be accessed via anonymous ftp at ftp://standard.pictel.com and ftp://ftp.imtc-files.org

I pixel and the position is denoted with i where $i = 1,\ldots I$. The pixel value in the original sequence in frame n and macroblock m at macroblock position i is denoted as $s_{n,m,i}$.

H.26L/JVT coding consists of the motion compensation and the residual coding stage. The task of residual coding is to refine signal parts that are not sufficiently well represented by motion-compensated prediction. From the viewpoint of bit allocation strategies, the various modes relate to various bit rate partitions. The concept of selecting appropriate coding options in many source-coding standards is based on rate-distortion based algorithms. The two cost terms "rate" and "distortion" are linearly combined and the mode is selected such that the total cost is minimized. This can be formalized by defining the set of selectable coding options for one macroblock as $\mathfrak{O}$. In hybrid video coding systems the macroblock mode can be selected from the set of macroblock modes $\mathfrak{M}$. In the following, we assume that we only transmit one I-picture at the beginning of the video sequence and P-pictures for the remainder. However, the presented algorithm can be extended easily to other picture types like B or multi-hypothesis pictures. Therefore, we assume that the set of macroblock modes consists of two subsets, one including macroblock modes, which employ temporal prediction, denoted as $\mathfrak{M}_{\mathrm{P}}$ and one including pure intra coding without any prediction denoted as $\mathfrak{M}_{\mathrm{I}}$. Obviously, for I-pictures the macroblock mode can only be selected from $\mathfrak{M}_{\mathrm{I}}$. In H.26L, not only the mode of the macroblock can be selected but also the reference frame from the set of accessible reference frames $\mathfrak{R}$ can be chosen [13]. The cardinality of set of reference frames $|\mathfrak{R}|$ specifies the maximum number of reference frames. The set of accessible coding options for P-frames is defined as all possible combinations of macroblock modes and reference frames, i.e. $\mathfrak{O} = \{\mathfrak{M}_{\mathrm{I}}, \mathfrak{M}_{\mathrm{P}} \times \mathfrak{R}\}$. Therefore, rate-constrained mode decision selects the coding option $o^*_{n,m}$ for macroblock m in frame n such that the Lagrangian cost functional is minimized, i.e.

$$o^*_{n,m} = \arg\min_{o \in \mathfrak{O}} \left(D_{n,m}(o) + \lambda R_{n,m}(o) \right). \qquad (1)$$

In the H.26L test model for coding efficiency, the distortion $D_{n,m}(o)$ is the sum of squared pixel differences (SSD), i.e.

$$D_{n,m}(o) = \sum_{i=1}^{I} \left| s_{n,m,i} - \hat{s}_{n,m,i}(o) \right|^2, \qquad (2)$$

where $\hat{s}_{n,m,i}(o)$ is the reconstructed pixel value at the decoder in frame n and MB m at position i when encoding with macroblock mode o. The rate $R_{n,m}(o)$ is simply obtained by encoding with mode o and the Lagrange parameter is selected as $\lambda = C_\lambda \cdot 2^{q/3}$ with $C_\lambda = 0.85$ [14].

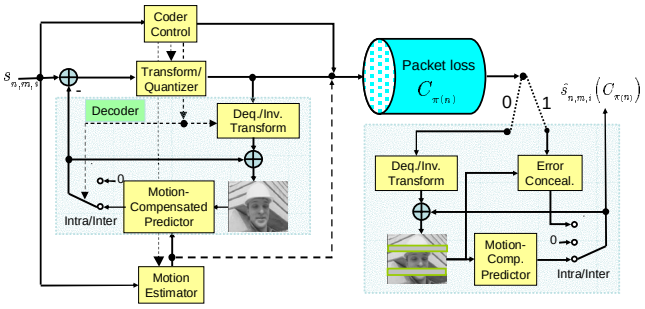

Figure 1 *Video transmission over packet loss channels.*

The generated video data is packetized and transmitted over a packet lossy channel. The channel behavior c when transmitting frame n is defined by a binary sequence $\{0,1\}^{\pi(n)}$ with $\pi(n)$ the number of packets necessary to transmit frame $1,\ldots,n$. A 0 in the channel sequence indicates a correctly received packet whereas a 1 indicates a lost packet. We denote the binary channel loss sequence up to frame n as $c_{\pi(n)}$ indicating the length of this sequence $\pi(n)$ in the index. We can assume that the decoder is aware of the channel behavior as appropriate error detection mechanisms like block check sequences and sequence numbering are applied in common transport protocols. Although feedback [16] might be used to convey the channel loss sequence to the encoder, in general we can not assume that the encoder is aware of the channel sequence $c_{\pi(n)}$ when encoding frame $n+1$. Therefore, we define the channel loss sequence as a random variable denoted as $C_{\pi(n)}$ and assume that the encoder knows at least the statistics of this sequence by, e.g.. RTCP messages.

The decoder decodes the received sequence of packets. Whereas correctly received packets are decoded as usual for the lost packet an error concealment algorithm has to be invoked. The reconstructed pixel $\hat{s}_{n,m,i}$ at position i in macroblock m and frame n depends on the encoding mode, the channel behavior and on the decoder error concealment. Note that due to the motion compensation process the reconstructed image depends not only on the lost packets for the current frame but in general on the entire channel loss sequence $C_{\pi(n)}$. We denote this dependency as $\hat{s}_{n,m,i}\left(o, C_{\pi(n)}\right)$. The encoder can get an estimate of the reconstructed value at the decoder and, therefore, of the expected distortion as

$$D_{n,m}\left(o, C_{\pi(n)}\right) = \sum_{i=1}^{I} E_{C_{\pi(n)}} \left| s_{n,m,i} - \hat{s}_{n,m,i}\left(o, C_{\pi(n)}\right) \right|^2, \qquad (3)$$

where the expectation is over the channel $C_{\pi(n)}$.

2.2 Estimate of Decoder Pixel Distortion

The estimate of the expected pixel distortion in packet loss environment has been addressed in several previous papers. Whereas for example in [11] or [12] models to estimate the pixel distortion are defined, the recursive optimal per-pixel estimate (ROPE) algorithm [10] provides an accurate estimation by keeping track of the first and second moment of $\hat{s}_{n,m,i}$, $E\{\hat{s}_{n,m,i}\}$ and $E\{\hat{s}^2_{n,m,i}\}$, respectively. However, the extension of the ROPE algorithm to H.26L/JVT coding is not straightforward. The in-loop filter, the sub-pel motion accuracy and the advanced error concealment require taking into account the expectation of products of pixels at different positions to obtain an accurate estimation which makes the ROPE unfeasible in this case. Therefore, we have chosen a different method for approximating the expected decoding distortion without attempting to provide a comparison of the complexity of the two methods which is subject to future work.

Let us assume that we have K copies of the random variable channel behavior at the encoder, denoted as $C_{\pi(n)}(k)$. Additionally, assume that the set of random variables $C_{\pi(n)}(k)$, $k = 1,\ldots,K$ are *identically* and *independently* distributed (*iid*). Then, as $K \to \infty$, it follows by the strong law of large numbers that

$$\frac{1}{K}\sum_{k=1}^{K} \left| s_{n,m,i} - \hat{s}_{n,m,i}\left(C_{\pi(n)}(k)\right) \right|^2 = E_{C_{\pi(n)}} \left| s_{n,m,i} - \hat{s}_{n,m,i}\left(C_{\pi(n)}\right) \right|^2, \qquad (4)$$

holds with probability 1. An interpretation of the left hand side leads to a simple solution of the previously stated problem to estimate the expected pixel distortion. In the encoder K copies

of the random variable channel behavior and the decoder are operated. The reconstruction of the pixel value depends on the channel behavior $C_{\pi(n)}(k)$ and the decoder including error concealment. The K copies of channel and decoder pairs in the encoder operate independently. Therefore, the expected distortion at the decoder can be estimated accurately in the encoder if K is chosen large enough.

2.3 Implementation of Mode Selection Algorithm

In [15] it was shown that the mode selection for packet lossy channels with packet loss probability p can be carried out according to

$$o^*_{n,m} = \arg\min_{o \in \mathfrak{O}} \left(\hat{D}_{n,m}(o) + \hat{C}_\lambda 2^{q/3} R_{n,m}(o) \right), \qquad (5)$$

where $\hat{D}_{n,m}(o)$ defines the expected distortion for macroblock m in frame n assuming that all transmission packets of frame n are received correctly, but the reference frames are erroneous based on the random packet loss sequence $C_{\pi(n-1)}$. Additionally, it was shown in [15] that the Lagrange parameter should be adapted to a value $\hat{C}_\lambda \le C_\lambda$ depending on the selected mode and the loss rate. However, the benefits are marginal and, therefore, due to simplicity it is proposed to set $\hat{C}_\lambda = C_\lambda$. For the error-robust macroblock mode and reference frame selection in frame n we therefore encode each macroblock m with each accessible macroblock mode $o \in \mathfrak{O}$. Then, for each combination (n, m, o) we decode this macroblock K times based on the reference frames generated by the independent channel realizations $C_{\pi(n-1)}(k)$ and compute the expected distortion $\hat{D}_{n,m}(o)$ similar to (4) as the arithmetic mean of the SSD for each decoded MB in each of the K channel-decoder pairs. After encoding the entire frame n the reference frame buffer in each channel-decoder pair is updated by independently applying channel realizations to each packet of the transmitted frame. Obviously, the method is rather complex as K times the complexity and the memory requirement of decoder is necessary in the encoder. However, due to the simplicity, robustness, and flexibility of the approach and the good converging properties for even low K this approach is very suitable to obtain performance bounds and obviously allows online encoding.

3 EXPERIMENTAL RESULTS

3.1 Estimated Decoder Distortion

We will compare the estimation of the decoder distortion for the algorithm developed in Section 2 to the ROPE algorithm [10] in the H.26L test model encoder. For the ROPE based system we use the nearest full-pel motion vector and we ignore the loop filter operation for the recursive estimation of first and second order of the pixel expectation. For both algorithms we apply simple previous frame concealment and we encode the test sequence *Foreman* (QCIF, 75 frames, 7.5 fps, QP 20, 11 MB per transmission packet, mode selection according to Section 2 with statistical independent packet loss rate 10%). In Figure 2 the PSNR for error-free transmission, the PSNR of the expected decoder distortion ($K = 500$ channel-decoder pairs) for each frame is plotted and compared to the corresponding ROPE algorithm. For most frame numbers the expected distortion and the ROPE give very similar results. However, for some parts the difference is significant. Especially in the last frames where sub-pixel global motion dominates the sequence activity, the full-pel ROPE can not accurately estimate the decoder distortion. This shows the necessity of the computation of the expected decoder according to Section 2 as the difference to the real distortion can be significant with full-pel ROPE. To obtain a reasonable value for the number of channel-decoder pairs in the encoder, the PSNR of the average distortion plus and minus the standard deviation of the distortion when using $K = 1$ and $K = 30$ channel-decoder pairs is shown. This gives hints on the performance of the decoder estimation when using a certain number of decoders. Note that the standard deviation from the actual expected distortion decays with $1/\sqrt{K}$. It can be observed that for $K = 30$ the expected distortion for most cases is very close to the real distortion as the standard deviation is small. Therefore, we select this value $K = 30$ for the implementation of the mode selection.

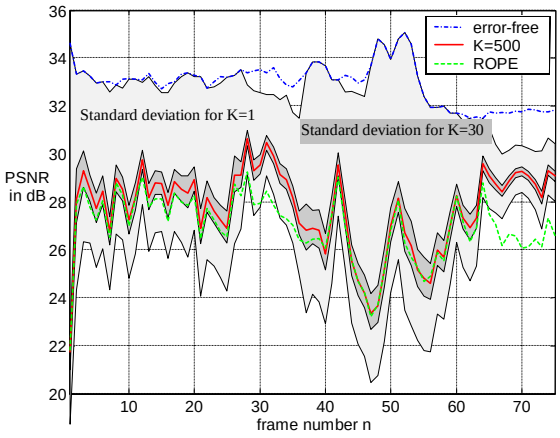

Figure 2 *PSNR of encoder and expected decoder distortion for K=500 and full-pel ROPE algorithm over frame number. Additionally, the PSNR of the average distortion plus and minus the standard deviation of the distortion when using* $K = 1$ *and* $K = 30$ *channel-decoder pairs is shown.*

3.2 Results for IP Common Test Conditions

We present results based on the common test conditions for wireline, conversational, IP/UDP/RTP based systems for H.26L/JVT coding and compare to H.263+ anchor performance [17]. The details of the simulation conditions are provided in [18]. However, we will summarize the most important simulation parameters. The test sequence *Foreman* (QCIF, 75 frames, 7.5 fps, maximum bit rate 64 kbit/s) and the test sequence *Paris* (CIF, 150 frames, 15 fps, maximum bit rate 144 kbit/s) are looped, encoded, packetized (2 RTP packets per video frame) and transmitted over Internet error patterns with different error patterns of approximately 3, 5, 10, and 20%. The performance measure applied is the modified average PSNR which compares the reproduced frame to all pictures of the source file at a frame rate of 30 fps to take into account lost frames. For H.263+, a rate control to meet the maximum bit-rate is used and an error resilient macroblock mode selection similar to the advanced method as described in [11] is applied. For H.26L a pseudo-random (PR) intra-update based on [7] with loss adaptive refresh rate as well as the rate-distortion optimized (RDO) mode selection according to Section 2 with $K = 30$ channel-decoder pairs with independent losses is used. The error concealment at the decoder is according to [19] whereas for the RD optimization the previous frame concealment is used. As the H.26L test

model does not yet include a rate control, one quantization parameter for the entire is selected in order to stay below the maximum bit-rate. Note that due slight differences in the simulation conditions, H.263+ and H.26L results are not exactly comparable. However, we are confident that the tendencies of the reported results hold. Also the statistics for the experiments are sufficient.

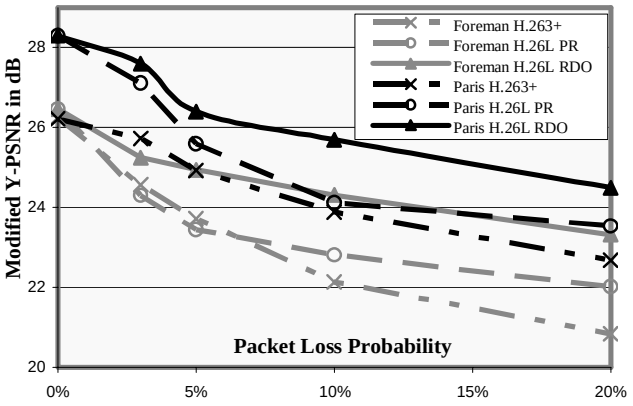

Figure 3 ***Modified Y-PSNR in dB over packet loss probability of Internet error patterns for H.263+ Internet anchor streams, H.26L with pseudo-random (PR) intra-updates and rate-distortion optimized (RDO) intra update.***

The results of several experiments are shown in **Figure 3**. The modified PSNR is plotted over the packet loss rate referring to the error patterns. It can be observed that H.26L outperforms H.263+ for both sequences even without the error-resilient RD-optimization as presented previously. The pseudo-random intra update is already quite effective. This is as the RD-optimization introduced for coding efficiency already selects intra MBs quite frequently to obtain good coding performance and the overall compression efficiency is significantly better. In addition, the RD optimization provides significant additional increase in quality. For example for the 10% error case the gains are about 2 dB in PSNR when compared with pseudo-random intra update. The subjective results match the results based on the modified PSNR very well. Appropriate sequences are presented.

4 CONCLUSIONS AND OUTLOOK

In this work we apply a widely accepted standard-compliant technique to enhance the quality of H.26l/JVT coded video transmitted over packet lossy networks. The macroblock mode and the reference frame selection are extended to include the expected decoder distortion in the Lagrangian mode decision This allows a careful placement of rate-expensive intra-coded macroblocks. A flexible and robust, but rather complex method to derive the expected decoder distortion at the encoder is introduced. This allows taking into account all H.26L/JVT features, any kind of channel statistics and the decoder error concealment in this mode selection process. The good performance of this new algorithm as well as the suitability of H.26L/JVT coding for packet lossy networks are verified for common Internet test conditions. Future work includes the combination of the work with advanced error concealment strategies [19], combination with appropriate rate control schemes, a comparison with an advanced ROPE algorithm to estimate the decoder distortion in the encoder, and, finally, the inclusions of feedback information in the mode selection.

5 REFERENCES

[1] J.K. Wolf, A.D. Wyner, and J. Ziv, “Source Coding for Multiple Descriptions”, Bell System Technical Journal, vol. 59, pp. 1417-1426, October 1980.

[2] T. Wiegand (ed.), “Committee Draft Number 1, Revision 0 (CD-1),” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-C167, May 2002.

[3] T. Stockhammer, M.M. Hannuksela, and T. Wiegand, “JVT/H.26L in 3G Wireless Environments”, submitted for IEEE CSVT, Special Issue on H.26L/JVT, April 2002.

[4] S. Wenger, “H.26L over IP: The IP-Network Adaptation Layer”, Proc. Packet Video Workshop 2002, Pittsburgh, PY, April 2002.

[5] Q. F. Zhu and L. Kerofsky, “Joint Source Coding, Transport Processing, and Error Concealment for H.323-Based Packet Video,” In Proceedings of the SPIE Conference on Visual Communications and Image Processing, volume 3653, pages 52–62, San Jose, CA, USA, January 1999.

[6] P. Haskell and D. Messerschmitt, “Resynchronization of Motion-Compensated Video Affected by ATM Cell Loss,” In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 3, pages 545–548, 1992.

[7] J. Liao and J. Villasenor, “Adaptive Intra Update for Video Coding over Noisy Channels,” In Proceedings of the IEEE International Conference on Image Processing, volume 3, pages 763–766, Lausanne, Switzerland, October 1996.

[8] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra. “Rate-Distortion Optimized Mode Selection for Very Low Bit Rate Video Coding and the Emerging H.263 Standard,” IEEE Transactions on Circuits and Systems for Video Technology, 6(2):182–190, April 1996.

[9] G.J. Sullivan and T. Wiegand, “Rate-Distortion Optimization for Video Compression,” IEEE Signal Processing Magazine, vol. 15, no. 6, pp. 74-90, Nov. 1998.

[10] R. Zhang, S. L. Regunathan, and K. Rose, “Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience,” in IEEE JSAC, vol. 18, no. 6, pp. 966-976.

[11] G. Cote, S. Shirani, F. Kossentini, “Optimal mode selection and synchronization for robust video communications over error-prone networks,” in IEEE JSAC, vol. 18, no. 6, pp. 952 -965.

[12] T. Wiegand, N. Färber, K. Stuhlmüller, and B. Girod, “Error-resilient video transmission using long-term memory motion-compensated prediction”, IEEE JSAC, vol. 18, no. 6, pp. 1050-1062.

[13] T. Wiegand and B. Girod, “Multi-frame Motion-Compensated Prediction for Video Transmission,” Kluwer Academic Publishers, Sept. 2001.

[14] T. Wiegand, and B. Girod, “Lagrangian Multiplier Selection in Hybrid Video Coder Control,” Proc. ICIP 2001, Thessaloniki, Greece, October 2001.

[15] T Stockhammer, D. Kontopodis, and T. Wiegand, “Rate-Distortion Optimization for JVT/H.26L Coding in Packet Loss Environment”, Proc. Packet Video Workshop 2002, Pittsburgh, PY, April 2002.

[16] B. Girod and N. Färber, “Feedback-Based Error Control for Mobile Video Transmission,” Proceedings of IEEE, vol. 97, no. 10, pp. 1707-1723, October 1999.

[17] S. Wenger, “Common Test Conditions for the H.323/Internet case,” ITU-T Standardization Sector Q15-K45, October 1999.

[18] S. Wenger and M. Horowitz, “Scattered Slices: Simulation Results,” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-C090, May 2002.

[19] V. Varsa, M.M. Hannuksela, and Y. Wang, “Non-normative error concealment algorithms,” ITU-T VCEG-N62, VCEG (SG16/Q6), Fourteenth Meeting, Santa Barbara, CA, September 2001.