ONE-PASS MULTI-LAYER RATE-DISTORTION OPTIMIZATION FOR QUALITY SCALABLE VIDEO CODING

Xiang Li[12]*, Peter Amon*[2]*, Andreas Hutter*[2]*, and André Kaup*[1]

[1]Chair of Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg, Erlangen, Germany
[2]Siemens Corporate Technology, Information & Communications, Munich, Germany

ABSTRACT

In this paper, a *one-pass* multi-layer rate-distortion optimization algorithm is proposed for quality scalable video coding. To improve the overall coding efficiency, the MB mode in the base layer is selected not only based on its rate-distortion performance relative to this layer but also according to its impact on the enhancement layer. Moreover, the optimization module for residues is also improved to benefit inter-layer prediction. Simulations show that the proposed algorithm outperforms the most recent SVC reference software. For eight test sequences, a gain of 0.35 dB on average and 0.75 dB at maximum is achieved at a cost of less than 8% increase of the total coding time.

Index Terms— H.264/AVC, SVC, Quality Scalable Video Coding, Multi-Layer RDO

1. INTRODUCTION

To support a diverse range of client capabilities and transmission channel capacities, the scalable video coding (SVC) extension [1] of H.264/AVC [2] was developed. Currently, this SVC extension is able to provide three kinds of scalability, i.e., temporal scalability, spatial scalability and quality scalability [3]. When compared with the scalable profiles in previous video coding standards, e.g., MPEG-4 Visual [4], the overall coding efficiency of this SVC extension has been greatly increased [3]. Nevertheless, there is much room for further improvement since in some cases the gap in rate-distortion performance to single layer coding is still significant [5].

In principle, the key technology distinguishing SVC from single layer coding is the inter-layer prediction, which is designed to remove the redundancies between layers. Intuitively, research should be focused on inter-layer prediction techniques in order to improve the coding efficiency of the SVC extension towards that of single layer coding.

In [6, 7], some new methods were proposed to improve the efficiency of inter-layer prediction in spatial scalability. Basically, their ideas are to increase the accuracy of the prediction by further exploring the correlation between layers. Although the gains by these algorithms are significant, their limitation is obvious. By assuming a fixed base layer, they try to independently optimize the enhancement layer and neglect the effect by the base layer. Intuitively, how to code the base layer will greatly influence the coding efficiency of the enhancement layer. Therefore, it is necessary to jointly optimize both base and enhancement layers.

Recently, a multi-layer rate-distortion optimization (RDO) algorithm was proposed [8]. By simultaneously considering both base and enhancement layers, the MB modes, motion vectors, and quantized residues are well selected to profit the inter-layer prediction. Simulations show that this algorithm is efficient. However, its computational complexity is very high since a *multi-pass* process has to be employed to enable the multi-layer optimization. Moreover, the quality scalability was not achieved by multiple layers but actually by a flexible sub-stream extracting method where a serious quality fluctuation may occur.

To improve the coding efficiency while keeping a reasonable computational complexity, a *one-pass* multi-layer rate-distortion optimization for quality scalable video coding using multiple layers is presented in this paper. Instead of checking the impact of the base layer on the enhancement layer after the real coding, the effect is estimated so that the *multi-pass* process in multi-layer RDO is avoided and so is the heavy computational payload. In addition, the optimization module for residues is also improved to profit inter-layer prediction. As will be shown by simulations, the proposed algorithm outperforms the most recent joint scalable video model JSVM 9.13.1 [9]. A gain of 0.35 dB on average and 0.75 dB at maximum is achieved for eight sequences at a cost of less than 8% increase of the total coding time.

The rest of the paper is organized as follows. First, the proposed multi-layer RDO algorithm is presented in Section 2. Then the performance of the algorithm is verified and discussed in Section 3. Finally, the whole paper is concluded in Section 4.

2. MULTI-LAYER RDO

In this section, the problem of multi-layer RDO is first formulated. Then the proposed two techniques, i.e., simplified MB

mode decision and improved optimization on residues are discussed, respectively.

2.1. Formulation of Multi-Layer RDO

In the most recent joint scalable video model JSVM 9.13.1 [9], each layer is optimized independently. For coding parameters $\mathbf{\Gamma_n^m}$ (including MB mode, motion vectors and quantized residues) of the MB m in layer n, define the rate-distortion (R-D) cost $C_n(\mathbf{\Gamma_n^m})$ as

$$C_n(\mathbf{\Gamma_n^m}) = D_n(\mathbf{\Gamma_n^m}) + \lambda_n \cdot R_n(\mathbf{\Gamma_n^m}), \tag{1}$$

where $D_n(\cdot)$ and $R_n(\cdot)$ denote the distortion and rate for the layer n, respectively, λ_n is the so-called Lagrange multiplier which is currently determined by quantization parameter [10] though a better performance can be expected by a more adaptive algorithm [11]. According to (2), the best $\mathbf{\Gamma_n^m}$ can be determined for the single layer coding.

$$\mathbf{\Gamma_n^m} = \arg\min_{\{\mathbf{\Gamma_n^m}\}} \{C_n(\mathbf{\Gamma_n^m})\}. \tag{2}$$

To count in the impact of the base layer coding on the enhancement layer, [8] proposed a conditional R-D cost for the enhancement layer in a two-layer optimization scenario,

$$\begin{aligned} C_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}) = {} & D_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}) \\ & + \lambda_{n+1}(R_n(\mathbf{\Gamma_n^m}) + R_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})), \end{aligned} \tag{3}$$

where subscript n and $(n+1)$ indicate the base and enhancement layer, respectively. Consequently, the best $\mathbf{\Gamma_n^m}$ is selected by the joint R-D cost

$$\mathbf{\Gamma_n^m} = \arg\min_{\{\mathbf{\Gamma_n^m}, \mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}\}} \{(1-w)C_n(\mathbf{\Gamma_n^m}) + w \cdot C_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})\}, \tag{4}$$

where $w \in [0,1]$ is a weighting factor which controls the trade-off between the optimizations for base and enhancement layers. On one hand, when w equals 0, the optimization reduces to the single layer RDO described in (2). On the other hand, w equaling 1 indicates the base layer is only optimized for the enhancement layer coding without considering the reconstruction quality of the base layer [8].

To reflect the accumulated rate for the enhancement layer, $R_n(\mathbf{\Gamma_n^m})$ is included in (3). However, we propose to discard this term. First, it makes (4) not symmetric. Intuitively, for the two extreme cases where $w = 0$ and $w = 1$, (4) should reduce to a single layer RDO. While with this term, (4) is still a multi-layer RDO since the rate for the base layer has to be considered even when w=1. Second, when extending (4) to a multi-layer scenario, counting the accumulated rate for all lower layers in RDO is less efficient since it will actually force a preference on the coding parameters with lower rate while sacrificing the quality. Therefore, we discard that term and revise the R-D cost for the enhancement layer as

$$\begin{aligned} C'_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}) = {} & D_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}) \\ & + \lambda_{n+1} \cdot R_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}). \end{aligned} \tag{5}$$

Replacing $C_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})$ in (4) with $C'_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})$, the proposed multi-layer RDO is formulated as

$$\mathbf{\Gamma_n^m} = \arg\min_{\{\mathbf{\Gamma_n^m}, \mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}\}} \{(1-w)C_n(\mathbf{\Gamma_n^m}) + w \cdot C'_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})\}. \tag{6}$$

Theoretically, (6) leads to an optimal solution. However, it is impractical to solve (6) since even for a two-layer case the product space by $\mathbf{\Gamma_n^m}$ and $\mathbf{\Gamma_{n+1}^m}$ is so huge that the corresponding computation payload is not affordable at all. Although a great simplification was achieved in [8], the computational complexity is still very high since a *multi-pass* is necessary for obtaining $C'_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m})$.

2.2. Simplified MB Mode Decision

To calculate the conditional cost in (5), a *multi-pass* coding process is needed. Intuitively, the computational complexity will be greatly reduced if the conditional cost can be simplified to a normal cost.

Currently in the SVC extension H.264/AVC, there are three inter-layer prediction techniques, i.e., inter-layer motion prediction (based on MB mode and motion information), inter-layer residual prediction (based on quantized residues) and inter-layer intra prediction (based on full reconstruction of the base layer) [3]. Since the full construction may be regarded as a special combination of motion prediction and residues, $\mathbf{\Gamma_n^m}$ can be replaced by the MB mode and motion information $\mathbf{M_n^m}$ plus quantized residues $\mathbf{r_n^m}$, namely

$$\mathbf{\Gamma_n^m} = (\mathbf{M_n^m}, \mathbf{r_n^m}). \tag{7}$$

In quality scalable video coding, it is reasonable to suppose that the best $\mathbf{M_{n+1}^m}$ is the same as the best $\mathbf{M_n^m}$, especially when the quantization gap between the two layers is not big. Therefore

$$\begin{aligned} \mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m} &= (\mathbf{M_{n+1}^m}, \mathbf{r_{n+1}^m})|(\mathbf{M_n^m}, \mathbf{r_n^m}) \\ &= (\mathbf{M_n^m}, \mathbf{r_{n+1}^m})|(\mathbf{M_n^m}, \mathbf{r_n^m}). \end{aligned} \tag{8}$$

Noticing that the residual prediction occurs in transform domain in quality scalable video coding, quantized residues $\mathbf{r_{n+1}}$ are derived as

$$\mathbf{r_{n+1}^m} = (\mathcal{T}(\mathbf{I^m} - \mathbf{P_{n+1}^m}) - \mathbf{r_n^m} \cdot Q_n)/Q_{n+1}, \tag{9}$$

where $\mathcal{T}(\cdot)$ is the transform operator, $\mathbf{I^m}$ represents the source MB, $\mathbf{P_{n+1}^m}$ denotes the inter or intra prediction for the MB m from the same layer, Q_n and Q_{n+1} indicate the quantization steps for the base and enhancement layers, and $\mathbf{r_n^m} \cdot Q_n$ implies the residual prediction from the reference layer.

According to the coding process, the reconstruction $\hat{\mathbf{I}}_\mathbf{n}^\mathbf{m}$ of the base layer is

$$\hat{\mathbf{I}}_\mathbf{n}^\mathbf{m} = \mathbf{P_n^m} + \mathcal{T}^{-1}(\mathbf{r_n^m}) \cdot Q_n, \tag{10}$$

where $\mathcal{T}^{-1}(\cdot)$ indicates the inverse transform. Considering the transform in the SVC extension of H.264/AVC is integer

DCT which is a linear transform, $\mathbf{r_n^m}$ can be derived from (10)

$$\mathbf{r_n^m} = \mathcal{T}(\hat{\mathbf{I}}_\mathbf{n}^\mathbf{m} - \mathbf{P_n^m})/Q_n. \quad (11)$$

Plugging (11) into (9),

$$\begin{aligned}\mathbf{r_{n+1}^m} &= (\mathcal{T}(\mathbf{I^m} - \hat{\mathbf{I}}_\mathbf{n}^\mathbf{m}) - \mathcal{T}(\mathbf{P_{n+1}^m} - \mathbf{P_n^m}))/Q_{n+1} \\ &\approx \mathcal{T}(\mathbf{I^m} - \hat{\mathbf{I}}_\mathbf{n}^\mathbf{m})/Q_{n+1},\end{aligned} \quad (12)$$

where $(\mathbf{P_{n+1}^m} - \mathbf{P_n^m})$ may approximately be regarded as zero since $\mathbf{P_{n+1}^m}$ should be similar to or slightly better than $\mathbf{P_n^m}$ in quality scalable video coding when the gap between Q_{n+1} and Q_n is not big. Consequently, (8) is further simplified by putting (12) into it, i.e.,

$$\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m} \approx (\mathbf{M_n^m}, \mathcal{T}(\mathbf{I^m} - \hat{\mathbf{I}}_\mathbf{n}^\mathbf{m})/Q_{n+1})|(\mathbf{M_n^m}, \mathbf{r_n^m}). \quad (13)$$

Note that terms representing the coding parameters for the layer $(n+1)$ disappear in the right hand of (13), which indicates that the dependency introduced by inter-layer prediction is broken. Accordingly, a *multi-pass* coding is not necessary any more.

Plugging (13) into (5), the R-D cost for the enhancement layer is derived as

$$\begin{aligned}&C'_{n+1}(\mathbf{\Gamma_{n+1}^m}|\mathbf{\Gamma_n^m}) \\ &\approx D_{n+1}((\mathbf{M_n^m}, \mathcal{T}(\mathbf{I^m} - \hat{\mathbf{I}}_\mathbf{n}^\mathbf{m})/Q_{n+1})|(\mathbf{M_n^m}, \mathbf{r_n^m})) \\ &+ \lambda_{n+1} \cdot R'_{n+1}((\mathbf{M_n^m}, \mathcal{T}(\mathbf{I^m} - \hat{\mathbf{I}}_\mathbf{n}^\mathbf{m})/Q_{n+1})|(\mathbf{M_n^m}, \mathbf{r_n^m})),\end{aligned} \quad (14)$$

where $R'_{n+1}(\cdot)$ denotes the residual rate in layer $(n+1)$. In fact, assuming that $\mathbf{M_{n+1}^m}$ is the same as $\mathbf{M_n^m}$ indicates that it can be perfectly predicted by inter-layer motion prediction. Therefore, only the part for residues should be counted for the rate of the enhancement layer.

Essentially, the proposed simplified MB mode decision method is formulated by putting (14) into (6), i.e., the MB mode with the minimal joint R-D cost should be selected as the best MB mode for the base layer. When extending (6) to a multi-layer scenario, a two-layer sliding window process is applied. That is for each layer, only its enhancement layer and itself are considered in (6) since simulations indicate that the correlation between two non-neighboring layers is normally quite weak so that the coding of the layer n will not show much effect on the layer $(n+2)$ and those above. After finishing one layer, the sliding window is moved upward by one layer and the optimized coding is conducted until all the layers are coded.

2.3. Improved Optimization on Residues

In both H.264/AVC reference software JM 14.1 [12] and SVC extension reference software JSVM 9.13.1 [9], there is a small module for the optimization on residues. The basic idea is to save bits by discarding small but expensive non-zero quantized residues. In practice, it checks the cost of each quantized residue where the cost is empirically predefined according to the amplitude and position of the quantized residue. If the accumulated cost for a 4x4 or 8x8 block is smaller than a certain threshold, the whole block will be discarded by forcing all residues to zero.

Generally, this optimization is very efficient for single layer coding. However, for multi-layer coding, its performance is not good since the residues may be used in inter-layer residual prediction. Discarding a whole block may reduce the rate for the base layer, but will also degrade the performance of the enhancement layer since zeros contribute nothing in prediction.

To improve the inter-layer residual prediction, some small residues should be kept from zeros. Here a simple but efficient method is proposed for the optimization on residues in multi-layer coding. That is, the threshold for discarding blocks is decreased by 3. By this approach, more small residues are kept to profit the residual prediction while the real "expensive" coefficients can be avoided as well.

Table 1. Simulation Results

sequences	w=0.25		w=0.5		$w = 0.75$	
	ΔP(dB)	ΔT	ΔP(dB)	ΔT	ΔP(dB)	ΔT
bus	0.23	6.87%	0.32	6.02%	0.29	5.69%
football	0.12	9.89%	0.23	10.33%	0.20	9.89%
foreman	0.25	6.28%	0.40	5.91%	0.44	4.79%
mobile	0.25	7.83%	0.34	7.83%	0.28	7.83%
city	0.26	7.61%	0.44	7.23%	0.49	7.04%
crew	0.27	7.92%	0.33	9.18%	0.29	9.34%
harbour	0.24	7.55%	0.34	7.55%	0.31	7.55%
soccer	0.24	6.00%	0.40	4.99%	0.42	4.33%
average	0.23	7.49%	0.35	7.38%	0.34	7.06%

3. SIMULATIONS AND DISCUSSIONS

The proposed algorithm was verified by the most recent joint scalable video model JSVM 9.13.1 [9]. Totally eight sequences defined in testing conditions for SVC coding efficiency [13] were coded on an Intel Xeon (X5355@2.66 GHz) PC with MS Windows Server 2003 R2 and 6 GB memory. To evaluate the algorithm in a similar quality range to that in [8], four quality layers were enabled with fixed quantization parameters, i.e QP =32, 30, 28, 26. All the frames of each sequence were encoded in IPPP structure (only one I frame at the very beginning). In addition, CABAC, fast search algorithm for motion estimation were enabled while temporal scalability, middle granularity scalability, 8x8 transform, and low complexity MB mode were disabled [9].

Table 1 summaries the simulation results for different w where ΔP denotes the average gain (calculated by [14]) in PSNR-Y over JSVM 9.13.1 for all layers of each sequence, ΔT represents the increase of the total coding time. In addition, three related R-D curves are drawn in Fig. 1. Similar to [8], w = 0.5 shows a best overall performance. Therefore, we focus on this case for the following discussions.

On average, a gain of 0.35 dB was achieved for eight se-

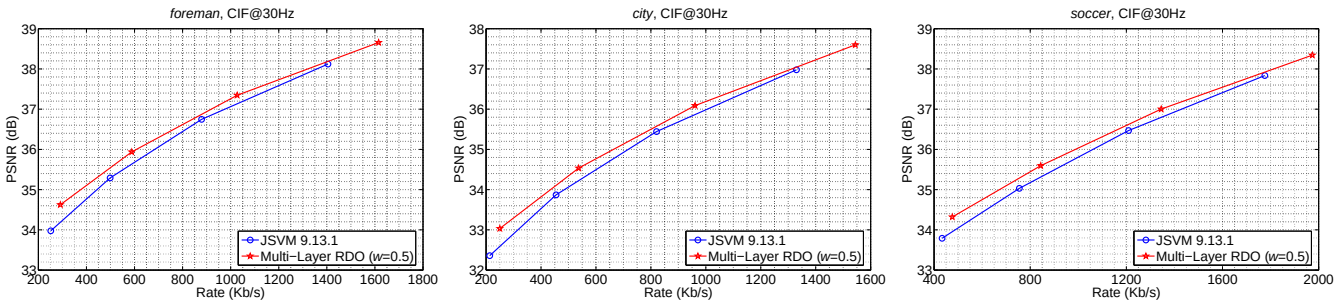

Fig. 1. Simulation Results. (Four quality layers were enabled where QP=32, 30, 28, 26)

quences at a cost of less than 8% increase of the total coding time. Moreover, except the 0.23 dB gain for *football*, the gains for other sequences are all above 0.3 dB, which indicates the proposed algorithm well adapts to different scenarios. Actually, more significant peak gains can be observed from Fig. 1. For *foreman* and *city*, about 0.75 dB gains can be noticed around 300 Kb/s and 250 Kb/s, respectively.

Unfortunately, it is difficult to fairly compare the coding efficiency of the proposed algorithm to [8]. As mentioned in Section 1, the quality scalability in [8] was not achieved by multiple quality layers. Instead, it was obtained by successively discarding quality enhancement representations of the B frames starting with the finest temporal level. Since only two layers were employed, no much layer overhead was introduced by layers so that the overall coding efficiency was close to that of the single layer coding. On the contrary, the proposed method is for the scenario with multiple layers where a higher overhead is introduced by more layers. Thus the efficiency gap between the proposed algorithm and the single layer coding is bigger. However, if we only examine the gains over JSVM software, the performance of the two algorithms are similar, i.e. both of them obtain a gain of 0.6 dB for *soccer* around 400 Kb/s.

In General, the proposed algorithm achieves a lower distortion at a cost of a slightly higher rate. This is because MB modes and residues with lower distortion are selected to benefit inter-layer prediction. Consequently, the overall coding efficiency is improved and a better performance over JSVM 9.13.1 is observed.

4. CONCLUSIONS AND FUTURE WORK

In this paper, a *one-pass* multi-layer rate-distortion optimization algorithm for quality scalable video coding is presented. To count in the impact of the base layer coding on the enhancement layer, the MB mode with minimal joint R-D cost is selected in the base layer so that the overall coding efficiency is improved. In addition, the optimization module for residues is also improved to benefit inter-layer residual prediction. Simulations verified that the proposed algorithm outperforms the most recent JSVM software. For eight sequences, a gain of 0.35 dB on average and 0.75 dB at maximum is achieved at a cost of less than 8% increase of the total coding time. For the next step, extending this work to spatial scalability is an interesting topic.

5. ACKNOWLEDGMENT

This work was achieved with help of the European Community's Seventh Framework Program through grant agreement ICT OPTIMIX nINFSO-ICT-214625.

6. REFERENCES

[1] G. J. Sullivan, T. Wiegand, and H. Schwarz, "ITU-T Rec. H.264 — ISO/IEC 14496-10 Advanced Video Coding Defect Report (JVT-Z210)," in *JVT Meeting (Joint Video Team of ISO/IEC MPEG & ITU-T VCEG)*, Antalya, Turkey, Jan. 2008.

[2] JVT, *Advanced Video Coding (AVC) - 3rd Edition*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 Part 10). 2004.

[3] M. Wien, H. Schwarz, and T. Oelbaum, "Performance Analysis of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1194–1203, 2007.

[4] ISO/IEC, *Coding of Audio-Visual Objects - Part 2: Visual*, ISO/IEC 14496-2 (MPEG-4 Part 2). 1999.

[5] 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, 2007.

[6] W. Yang, G. Rath, and C. Guillemot, "Scalable video coding with interlayer signal decorrelation techniques," *EURASIP J. Advances in Signal Process.*, 2007.

[7] R. Xiong, J. Xu, and F. Wu, "In-scale motion compensation for spatially scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 2, pp. 145–158, 2008.

[8] H. Schwarz and T. Wiegand, "R-D Optimized Multi-Layer Encoder Control for SVC," in *IEEE Int. Conf. on Image Process. (ICIP)*, 2007, pp. II– 281–II – 284.

[9] JVT, "H.264/SVC reference software (JSVM 9.13.1) and manual," CVS sever at garcon.ient.rwth-aachen.de, Jul. 2008.

[10] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *IEEE Int. Conf. on Image Process. (ICIP)*, Thessaloniki, Greece, 2001, pp. 542–545, vol.3.

[11] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Extended Lagrange multiplier selection for hybrid video coding using interframe correlation," in *Picture Coding Symposium (PCS)*, Lisbon, Portugal, Nov. 2007.

[12] JVT, "H.264/AVC reference software (JM14.1)," http://iphome.hhi.de/suehring/tml/, Jul. 2008.

[13] M. Wien and H. Schwarz, "Testing Conditions for SVC Coding Efficiency and JSVM Performance Evaluation (JVT-Q205)," in *JVT Meeting (Joint Video Team of ISO/IEC MPEG & ITU-T VCEG)*, 2005.

[14] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves (VCEG-M33)," in *VCEG Meeting (ITU-T SG16 Q.6)*, Austin, Texas, USA, Apr. 2001.