Videocodierung mit mehreren Referenzbildern

Video Coding with Multiple Reference Frames

Bernd Girod, Markus Flierl, Stanford University, U.S.A.

Zusammenfassung Bewegungskompensierte Prädiktion mit Hilfe mehrerer vergangener oder zukünftiger Referenzbilder, so genannte Multiframe-Prädiktion, kann die Kompression von Videosignalen verbessern. Multiframe-Prädiktion lässt sich in Verbindung mit den von MPEG bekannten P-Bildern, oder auch – durch Multihypothesen-Prädiktion – in Verbindung mit B-Bildern verwenden. Dieser Aufsatz fasst die aktuellen Entwicklungen zusammen, die inzwischen weitgehend in die Videocodier-Standards ITU-T Rec. H.263 und H.264/AVC – dem neuen gemeinsamen Standard von ITU-T VCEG und ISO/IEC MPEG – übernommen wurden. H.264/AVC Video übertrifft MPEG-2 bereits jetzt um mehr als den Faktor 2X. ►►►

Summary Motion-compensated prediction based on multiple previous or future frames can enhance the compression efficiency of video coding. Multiframe prediction can be applied as an extension to MPEG P-Pictures, but also to B-Pictures in the form of multihypothesis prediction. We review recent advances, several of which have been embraced by the ITU-T Rec. H.263 and the emerging H.264/AVC standard, jointly developed by ITU-T VCEG and ISO/IEC MPEG. Already, H.264/AVC video outperforms MPEG-2 by a factor greater than 2X.

KEYWORDS I.4.2 [Image Processing]: Compression, Computing Methodologies, Image Processing, Video Coding, Motion Compensation

1 Einleitung

Der Durchschnittshaushalt in den Vereinigten Staaten von Amerika konsumiert jährlich mehr als 1500 Fernsehstunden. Bisher wird diese Nachfrage nur teilweise mit digitalen Übertragungsverfahren gedeckt. Wenn wir dennoch einmal annehmen, es würden ausschließlich digitale Systeme mit MPEG-2 Codierung eingesetzt, dann entspricht dem oben genannten Konsum ein jährliches Datenvolumen von etwa 3.000 Gigabyte (oder 3 Terabyte) pro Haushalt [1]. Multipliziert mit den 70 Millionen US-Haushalten kumuliert sich die Datenrate auf mehr als 200 Exabyte pro Jahr (1 Exabyte = 10^{18} Byte). Vergleicht man dies mit der jährlichen weltweiten Produktion neuen Materials (geschätzt auf 1–2 Exabyte [1]) oder mit dem gesamten jährlichen Internet-Backbone-Verkehr (in den Vereinigten Staaten von Amerika derzeit ebenfalls 1 Exabyte mit einem jährlichen Wachstumsfaktor von zwei [2]), so ist diese Zahl beeindruckend und verdeutlicht die Bedeutung effizienter Videocodierung.

Derzeit verwenden digitale Kabel- und Satellitensysteme den ISO/IEC Standard MPEG-2 [3] zur Codierung von Videosignalen. Moderne Decoder-Chips sind normalerweise so leistungsstark, dass sie gleichzeitig mehrere *Standard-Definition Television* (SDTV) Kanäle decodieren können. Bereits heute wird im Endgerät neben dem leistungsstarken Decoder oft noch ein Encoder zur Kompression und lokalen Speicherung von Programmen hinzugefügt. Eine Weiterentwicklung dieser Systeme ist absehbar: Zukünftige digitale Empfänger werden einige Tausend Stunden Programm aus einer Vielzahl von verschiedenen Quellen speichern können, einschließlich digitaler Programme über Kabel und Satellit als auch *Video-On-Demand* mittels Internet. Zusatzfunktionen wie etwa Videokonferenz werden ebenfalls unterstützt werden. Mit zunehmender Vernetzung im privaten Haushalt ist ein weiterer Integrationsschritt vorstellbar: Digitale Empfänger entwickeln sich zu *Media Gateways*, die mehrere Audio- und Video-Endgeräte über ein *Local Area Network* (LAN) mit verdrahteten und drahtlosen Segmenten versorgen können.

Für all diese neuen Anwendungen ist eine hocheffiziente Co-

dierung wichtig. Das einführende Zahlenbeispiel macht deutlich, dass selbst eine moderate Verbesserung der Kompressionseffizienz um 10% den kumulierten Datenkonsum der US-Haushalte um das 20-fache des derzeitigen Internet-Backbone-Verkehrs reduziert. Seit 1996, dem Jahr der Standardisierung von MPEG-2, wurden wesentliche weitere Fortschritte in der Videocodierung erzielt. Einer der weitreichendsten ist das Konzept der bewegungskompensierten Multiframe-Prädiktion. Es wurde zuerst als Ergänzung des ITU-T Standards H.263 [4] standardisiert und ist von Anfang an wichtiger Bestandteil des neuen H.264/AVC Standard [5]. Dieser neue Standard wird gemeinsam von der *ITU-T Video Coding Experts Group* (VCEG) und der *ISO/IEC Motion Picture Experts Group* (MPEG) entwickelt.

Der vorliegende Beitrag konzentriert sich auf die neuen Möglichkeiten der Videocodierung mit mehreren Referenzbildern. Abschnitt 2 präsentiert das Konzept der Multiframe-Prädiktion; es handelt sich dabei um eine Erweiterung der derzeitigen P-Bilder. Abschnitt 3 erläutert die Multihypothesen-Prädiktion – eine Erweiterung der derzeitigen B-Bilder. Abschnitt 4 vergleicht schließlich, anhand einer Videotestsequenz, die Leistungsfähigkeit des neuen H.264/AVC Standard mit MPEG-2.

2 Multiframe-Prädiktion

Vor etwa 20 Jahren wurden Intra-Bild-Verfahren zu Inter-Bild-Verfahren weiterentwickelt, um die Datenraten bei der Videocodierung zu verringern. Die Gewinne durch diesen radikalen Schritt waren zu Beginn noch nicht sehr beeindruckend. Mit der Zeit jedoch wurde die Inter-Bild-Kompression bedeutend verbessert. Die benötigte Speicher- und Rechenkapazität ist allerdings im Vergleich zu Intra-Bild-Verfahren um zwei Größenordnungen höher. Durch die kontinuierlich fallenden Kosten für Halbleiter kann heute ein weiterer Entwicklungsschritt vollzogen werden, durch den sich die Komplexität noch einmal wesentlich erhöht.

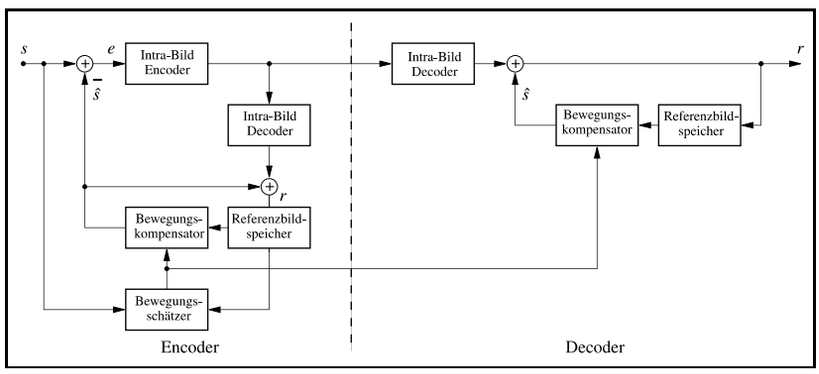

Bild 1 Hybrider Video-Codec mit Bewegungskompensation und mehreren Referenzbildern. Der Bewegungskompensator erzeugt mit Hilfe der Referenzbilder ein Prädiktionsbild $\hat{s}$. Beim Encoder codiert der Intra-Bild Encoder nur das Prädiktionsfehlerbild e, das der Differenz zwischen Originalbild s und Prädiktionsbild $\hat{s}$ entspricht. Beim Decoder wird zu dem decodierten Prädiktionsfehlerbild das Prädiktionsbild $\hat{s}$ addiert und dadurch das rekonstruierte Bild r erzeugt. Die Parameter der Bewegungskompensation werden vom Bewegungsschätzer des Encoders bestimmt und als Seiteninformation übertragen.

Bewegungskompensierte Multiframe-Prädiktion erweitert in blockbasierten hybriden Video-Codecs den Bewegungsvektor um einen Bildreferenz-Parameter. Dieser Parameter erlaubt die Verwendung mehrerer Referenzbilder im Vergleich zur herkömmlichen bewegungskompensierten Prädiktion, die nur das zuletzt decodierte Bild heranzieht [6]. Der Referenzbildspeicher im Encoder und Decoder behält Bilder, die vorteilhaft für die bewegungskompensierte Prädiktion sind (Bild 1). In den meisten Fällen verbessert die bewegungskompensierte Multiframe-Prädiktion die Kompressionseffizienz. Da der Bildreferenz-Parameter als zusätzliche Seiteninformation übertragen werden muss, ist eine intelligente Kontrolle der Datenrate notwendig. Diese wird erreicht, indem die besten Parameter der Bewegungskompensation unter Berücksichtigung einer Ratennebenbedingung bestimmt werden.

Bild 2 veranschaulicht am Beispiel des neuen Standards H.264/AVC die Kompressionseffizienz der Multiframe-Prädiktion. Die durch Prädiktion mit einem vorangehenden Referenzbild erzielte Bitrate wird mit der Bitrate für Multiframe-Prädiktion mit 5 vorangehenden Referenzbildern verglichen. Mit einer Speichergröße von 5 Bildern kann die Datenrate bei gleicher Bildqualität für die Videosequenz *Mobile & Calendar* in CIF Auflösung um 12% reduziert werden. Bei H.264/AVC

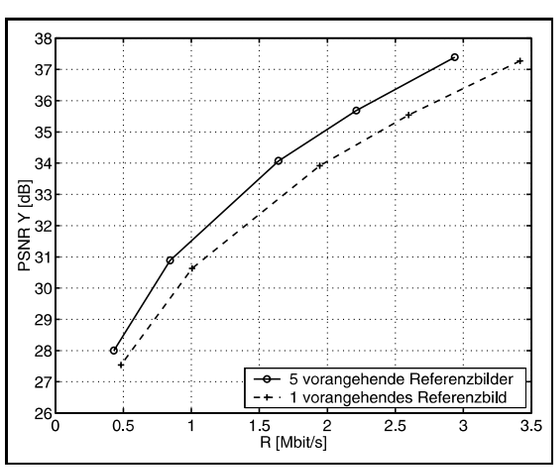

Bild 2 PSNR der Luminanz-Komponente über der Datenrate für die Videosequenz *Mobile & Calendar* in CIF Auflösung und 30 Bilder pro Sekunde, komprimiert mit H.264/AVC. Die Qualität der P-Bilder mit nur einem vorangehenden Referenzbild wird verglichen mit der von P-Bildern mit 5 vorangehenden Referenzbildern.

werden typisch die 5 vorangehenden Referenzbilder verwendet. Prinzipiell können beliebig viele Referenzbilder für die Prädiktion verwendet werden, jedoch ist im Allgemeinen eine Sättigung der Effizienzsteigerung zu beobachtet. Diese typische Anzahl bietet einen guten Kompromiss hinsichtlich der Komplexität.

Ursprünglich wurde bewegungskompensierte Multiframe-Prädiktion als ein Verfahren zur Verbesserung der Fehlerrobustheit von codiertem Video vorgeschlagen. Frühe Arbeiten von Budagavi und Gibson [7] verwenden einen zufällig variierenden Bildreferenz-Parameter und Annex N des Standards H.263 [8] erlaubt eine adaptive Referenzbildauswahl in Abhängigkeit der fehlerfrei decodierten Referenzbilder. Die Verbesserung der Kompressionseffizienz mit bewegungskompensierter Multiframe-Prädiktion wurde zuerst 1997 von Wiegand, Zhang und Girod vorgeschlagen [6; 9]. Die Verfahren wurden dann 1999 Bestandteil von H.263 Annex U „Enhanced Reference Picture Selection" [4] und sind jetzt ein integraler Bestandteil des neuen Standards H.264/AVC.

3 Multihypothesen-Prädiktion

B-Bilder sind Bilder in einer Bewegtbildsequenz, die unter Verwendung von zurückliegenden und zukünftigen Referenzbildern codiert werden. Eine Linearkombination aus vorwärts- und rückwärtsprädizierten Signalen ermöglicht die so genannte „bi-direktionale Prädiktion" (Bild 3). Eine solche lineare Überlagerung ist jedoch nicht notwendigerweise auf vorwärts- und rückwärts-prädizierte Signale beschränkt [10; 11]. Multihypothesen-Prädiktion, wie in [12] vorgeschlagen, erlaubt auch eine allgemeinere Form von B-Bildern [13; 14].

Für die bi-direktionale Prädiktion ist eine unabhängige Schätzung der vorwärts- und rückwärts-prädizierten Signale praktikabel, obwohl

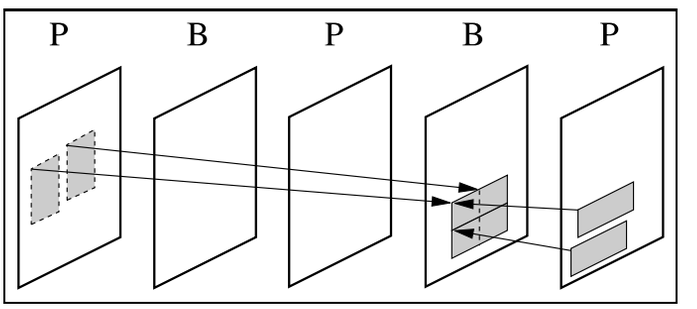

Bild 3 Bi-direktionale Prädiktion erlaubt nur eine Linearkombination eines zurückliegenden und eines zukünftigen Prädiktionssignals.

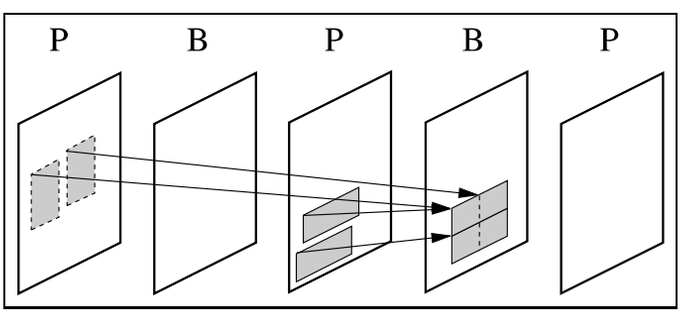

Bild 4 Im Gegensatz zur bi-direktionalen Prädiktion erlaubt die Multihypothesen-Prädiktion auch eine Linearkombination zweier zurückliegender Prädiktionssignale.

durch eine gemeinsame Schätzung die Kompressionseffizienz noch verbessert werden kann. Dagegen ist bei der Multihypothesen-Prädiktion eine gemeinsame Schätzung im Allgemeinen notwendig [15]. Eine unabhängige Schätzung könnte die Effizienz in diesem Fall sogar verschlechtern.

Multihypothesen-Prädiktion hebt die Einschränkung der bi-direktionalen Prädiktion auf, die nur eine Linearkombination von Paaren mit vorwärts- und rückwärts-prädizierten Signalen erlaubt. Zusätzliche Kombinationen wie (vorwärts, vorwärts) und (rückwärts, rückwärts) ergeben sich durch Erweiterung des uni-direktionalen Bildreferenz-Parameters zu einem bi-direktionalen Bildreferenz-Parameter (Bild 4). Jedes einzelne der überlagerten Prädiktionssignale kann als eine „Hypothese" für das aktuelle Bild interpretiert werden, deshalb der Name „Multihypothesen-Prädiktion". Die Überlagerung von zwei Prädiktionssignalen ist strenggenommen eine „Zweihypothesen-Prädiktion", aber dieser Ausdruck wird nicht verwendet.

Multihypothesen-Prädiktion schließt den Fall der bi-direktionalen Prädiktion ein, wenn die erste Hypothese vom zurückliegenden, die zweite Hypothese vom zukünftigen Referenzbild gewählt wird. In anderen Worten, bi-direktionale Prädiktion schränkt die Menge der möglichen Referenzbildpaare ein. Es ist nicht überraschend, dass eine größere Menge an möglichen Referenzbildpaaren die Kompressionseffizienz der B-Bilder verbessert.

Bild 5 vergleicht die Videoqualität der verallgemeinerten B-Bilder mit der der klassischen B-Bilder bei gegebener Datenrate für die Videosequenz *Mobile & Calendar* in CIF Auflösung. Die verallgemeinerten B-Bilder verwenden bewegungskompensierte Prädiktion mit bis zu zwei Hypothesen. Diese werden von 5 zurückliegenden und 3 zukünftigen Referenzbildern (P-Bilder) ausgewählt. Die klassischen B-Bilder basieren auf bi-direktionaler Prädiktion und verwenden als Referenz nur das jeweils vorangehende und folgende P-Bild. Für dieses Experiment werden jeweils zwei B-Bil-

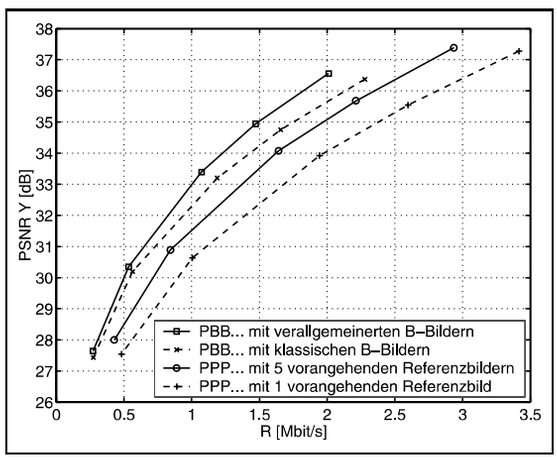

Bild 5 PSNR der Luminanz-Komponente über der Datenrate für die Videosequenz *Mobile & Calendar* in CIF Auflösung und 30 Bilder pro Sekunde, komprimiert mit H.264/AVC. Zwei verallgemeinerte B-Bilder werden von P-Bildern eingeschlossen. Verallgemeinerte B-Bilder mit bis zu zwei Hypothesen und mehreren Referenzbildern sind effizienter als klassische B-Bilder mit bi-direktionaler Prädiktion vom jeweils vorangehenden und folgenden P-Bild.

der zwischen aufeinanderfolgende P-Bilder eingefügt. Bild 5 veranschaulicht weiterhin einen Vergleich zur kausalen Codierung mit P-Bildern und variierender Anzahl an Referenzbildern. Es ist zu beobachten, dass die anti-kausale Codierung mit B-Bildern effizienter ist. Allerdings ist die Codierverzögerung größer als bei Prädiktion mit Referenzbildern ausschließlich aus der Vergangenheit.

Zusammenfassend lässt sich festhalten, dass verallgemeinerte B-Bilder eine konzeptionelle Trennung zwischen der Auswahl an Referenzbildern und der linearen Überlagerung von Prädiktionssignalen zulassen. Zum Beispiel können verallgemeinerte B-Bilder mit ausschließlicher Vorwärts-Prädiktion wie P-Bilder eingesetzt werden. Der Vorteil liegt dabei in der Effizienzsteigerung durch linear überlagerte Prädiktionssignale ohne zusätzliche Verzögerung durch anti-kausale Referenzbilder. Die höchste Kompression wird erreicht, wenn Multihypothesen-Prädiktion mit vergangenen und zukünftigen Referenzbildern eingesetzt wird.

4 Vergleich zwischen H.264/AVC und MPEG-2

MPEG-2 ist derzeit der vorherrschende Videokompressionsstandard für digitale Fernsehsignale. Ein Leistungsvergleich mit dem neuen H.264/AVC Standard ist daher aufschlussreich und wurde von der ITU-T VCEG angestellt. Bild 6 stellt beispielhaft ein Ergebnis dar.

Im Unterschied zu den vorangehenden Experimenten in diesem Beitrag liegt die Videosequenz *Mobile & Calendar* im zeilenverschränkten Format vor. Abgesehen von DVD Filmen liegen fast alle Fernsehsignale heute noch mit Zeilenverschränkung vor. Der Standard MPEG-2 wurde deshalb auch besonders für das zeilenverschränkte Format optimiert. Zum Zeitpunkt der Ergebnisse in Bild 6 war H.264/AVC noch nicht vollständig für dieses Format optimiert. Dennoch wurde im Vergleich mit MPEG-2 bei gleichem PSNR eine um 60% geringere Datenrate erzielt [16]. Die Ergebnisse mit H.264/AVC wurden durch Codierung der Halbbilder erzielt [17]. Weitere Verbesserungen können durch adaptive Auswahl zwischen Halb- und Vollbildern auf Bild- oder Makroblock-Ebene erzielt werden. Die gezeigten Ergebnisse schließen für H.264/AVC Multiframe-Prädiktion mit ein, aber noch keine verallgemeinerten B-Bilder.

5 Zusammenfassung

Die Bewegungskompensation mit mehreren Referenzbildern verbessert die Effizienz der Videocodierung und wird zunehmend in moderne Videokompressionsstandards integriert. Wenn neue Verfahren wie dieses mehr Speicher- und Rechenleistung benötigen, dann sind Bedenken der Hardwarehersteller vorhersehbar. Wie immer wird sich aber langfristig *Moore's Law* durchsetzen und der Kompressionseffizienz wird Priorität eingeräumt werden. Gary Sullivan, Raporteur der H.264/AVC Standardisierung, formulierte diese Entwicklung folgendermaßen: „*What once seemed like a strange and wasteful idea of requiring storage and searching of extra old pictures is becoming accepted practice – indeed it is the previous practice of throwing away the old decoded picture that has started to seem wasteful.*“ [18]

Verfahren mit mehreren Referenzbildern werden wahrscheinlich zuerst in Systemen und Endgeräten verwendet, die nicht an MPEG-2 gebunden sind und aus effizienterer Kompression Nutzen ziehen. Als Beispiele sind Internet-Video-Streaming, lokale Programmspeicherung und Multimedia-Vernetzung im privaten Haushalt zu nennen. Sobald

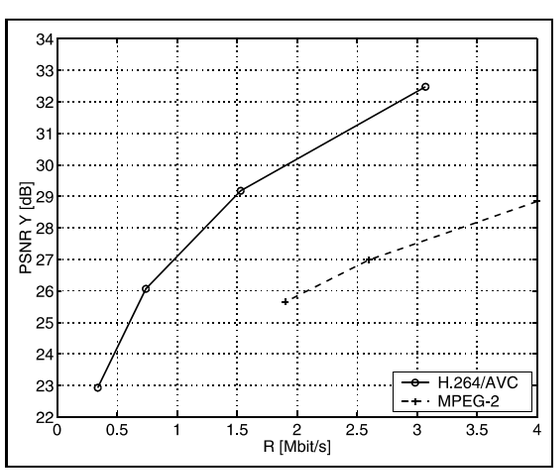

Bild 6 PSNR der Luminanz-Komponente über der Datenrate für die zeilenverschränkte Videosequenz *Mobile & Calendar* (352×480), komprimiert mit H.264/AVC und MPEG-2. Die nummerischen Werte wurden den Dokumenten [16] und [17] entnommen.

digitale Endgeräte mit dem neuen Format kompatibel sind, ist es naheliegend, die verfügbaren Rundfunk-Kapazitäten besser zu nutzen und H.264/AVC an Stelle von MPEG-2 auch für das digitale Fernsehen zu verwenden.

Danksagung

Die Arbeiten zu dem vorliegenden Beitrag wurden vom Graduiertenkolleg „Dreidimensionale Bild-Analyse und -Synthese“ der Universität Erlangen-Nürnberg gefördert und vom *Information Systems Laboratory* der *Stanford University* unterstützt.

Literatur

[1] P. Lyman, H.R. Varian: „How much information“, 2000. Retrieved from http://www.sims.berkeley.edu/how-much-info.

[2] K.G. Coffman, A.M. Odlyzko: „Internet growth: Is there a „Moore's Law“ for data traffic“, 2001. Retrieved from http://www.research.att.com/~amo/doc/internet.moore.pdf.

[3] ISO/IEC: *13818-2 Information Technology - Generic Coding of Moving Pictures and Associated Audio Information: Video (MPEG-2)*, 1996.

[4] ITU-T: *Recommendation H.263 (Video Coding for Low Bitrate Communication) Annex U*, 2000.

[5] ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group: *Study of Final Committee Draft of Joint Video Specification (ITU-T Rec. H.264, ISO/IEC 14496-10 AVC)*, Mar. 2003, ftp://ftp.imtc-files.org/jvt-experts/2003_03_Pattaya/JVT-G050d4.zip.

[6] T. Wiegand, X. Zhang, B. Girod: „Long-term memory motion-compensated prediction,“ *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, Feb. 1999.

[7] M. Budagavi, J.D. Gibson: „Multiframe block motion compensated video coding for wireless channels,“ in *Thirtieth Asilomar Conf. on Signals, Systems and Computers*, Nov. 1996, vol. 2, pp. 953–957.

[8] ITU-T: *Recommendation H.263, Version 2 (Video Coding for Low Bitrate Communication)*, 1998.

[9] T. Wiegand, X. Zhang, B. Girod: „Block-based hybrid video coding using motion-compensated long-term memory prediction,“ in *Proc. of the Picture Coding Symposium*, Berlin, Germany, Sept. 1997, pp. 153–158.

[10] M. Flierl, T. Wiegand, B. Girod: „Rate-constrained multihypothesis prediction for motion compensated video compression,“ *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, pp. 957–969, Nov. 2002.

[11] M. Flierl, T. Wiegand, B. Girod: „Multihypothesis pictures for H.26L,“ in *Proc. of the IEEE International Conf. on Image Processing*, Thessaloniki, Greece, Oct. 2001, vol. 3, pp. 526–529.

[12] M. Flierl, B. Girod: „Multihypothesis predition for B frames,“ Document VCEG-N40, ITU-T Video Coding Experts Group, Sept. 2001, http://standards.pictel.com/ftp/video-site/0109_San/VCEG-N40.doc.

[13] M. Flierl, B. Girod: „Generalized B pictures,“ in *Proc. of the Workshop on MPEG-4*, San Jose, CA, June 2002.

[14] M. Flierl, B. Girod: „Generalized B pictures and the draft H.264/AVC video compression standard,“ *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 587–597, July 2003, invited paper.

[15] M. Flierl, B. Girod: „Multihypothesis motion estimation for video coding,“ in *Proc. of the Data Compression Conference*, Snowbird, Utah, Mar. 2001, pp. 341–350.

[16] P. Borgwardt: *Handling Interlaced Video in H.26L*, ITU-T Video Coding Experts Group, Sept. 2001, http://standards.pictel.com/ftp/video-site/0109_San/VCEG-N57.doc.

[17] M. Gallant, L. Winger, G. Côté: *Interlaced Field Coding Core Experiment*, ITU-T Video Coding Experts Group, Dec. 2001, http://standards.pictel.com/ftp/video-site/0112_Pat/VCEG-O40.doc.

[18] T. Wiegand, B. Girod: *Multi-Frame Motion-Compensated Prediction for Video Transmission*, Kluwer Academic Publishers, 2001.

1

2

1 Prof. Dr.-Ing. Bernd Girod ist Professor der Elektrotechnik am *Information Systems Laboratory* der *Stanford University, CA, USA*. Seine Forschungsinteressen umfassen vernetzte Multimediasysteme, Videokompression sowie dreidimensionale Bild-Analyse und -Synthese. Er ist Verfasser von mehr als 250 Fachveröffentlichungen und mehreren Büchern zu diesen Themen. Von 1993 bis 1999 war er Ordinarius für Nachrichtentechnik an der Universität Erlangen-Nürnberg. Als Entrepreneur hat Girod erfolgreich mit einer Reihe von Gründungsfirmen gearbeitet, darunter Polycom (Nasdaq: PLCM) Vivo Software, 8x8 (Nasdaq: EGHT), und RealNetworks (Nasdaq: RNWK). Er promovierte zum Dr.-Ing. an der Universität Hannover und hat einen Master-Grad vom Georgia Institute of Technology in Atlanta. Professor Girod ist ein „Fellow of the IEEE“ und „2002 Distinguished Lecturer“ der IEEE Signal Processing Society. Die European Signal Processing Society verlieh ihm 2002 den Preis für die beste Veröffentlichung.
Adresse: Stanford University, Information Systems Laboratory, Stanford, CA 94305-9510, USA. E-Mail: bgirod@stanford.edu

2 Dipl.-Ing. Markus Flierl studierte Elektrotechnik an der Universität Erlangen-Nürnberg und erreichte 1997 seinen Abschluss als Diplomingenieur. Von 1999 bis 2001 war er Stipendiat des Graduiertenkollegs der Universität Erlangen-Nürnberg. Am *Information Systems Laboratory* der *Stanford University* war er bis Ende 2002 *Visiting Researcher*. Seine Interessen umfassen die Themen Bewegung in Bildsequenzen, multidimensionale Signalverarbeitung und Datenkompression.
Adresse: E-Mail: mflierl@ieee.org