P. GERKEN

Quellenmodell des MPEG-4-Video-Verifikationsmodells und weiterführende Quellenmodelle

Das Quellenmodell, das dem MPEG-4-Video-Verifikationsmodell zugrunde liegt, wird diskutiert und mit den Quellenmodellen bisheriger Standards [2, 8, 9, 11] verglichen. Ausgehend hiervon werden Quellenmodelle weiterführender Codiertechniken, die bei der Videocodierung mit sehr niedrigen Übertragungsbitraten eingesetzt werden können, analysiert. Hierzu zählen die Quellenmodelle, die der 3D-objektbasierten Codierung, der wissensbasierten Codierung sowie der semantischen Codierung zugrunde liegen. Ein Beispiel für einen mehrstufigen Coder wird aufgeführt, der auf verschiedenen Quellenmodellen basierenden Codiertechniken vereint.

... Peter Gerken ist ..... am Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung der Universität Hannover.

1. Einleitung

Codierverfahren lassen sich mit Hilfe der Quellenmodelle beschreiben, auf denen sie basieren. Solche Quellenmodelle stellen eine mehr oder weniger genaue Beschreibung der Realität dar. Wird durch die Modellannahmen ein reales Signal exakt beschrieben, so läßt es sich fehlerfrei mit Hilfe der Modellparameter beschreiben. In diesem Fall ist lediglich eine Codierung und Übertragung der Modellparameter notwendig, nicht jedoch des Signals selbst. Im allgemeinen stellen jedoch die Quellenmodelle nur grobe Näherungen der Realität dar, so daß die Codierung eines Residuums, das sich aus der Differenz von realem und modelliertem Signal ergibt, erforderlich ist. Je nach Interpretation kann man dieses Residuum als Modellierungs- oder Prädiktionsfehler bezeichnen.

Ein Quellenmodell besteht aus einem Objektmodell, einem Bewegungsmodell, einem Kameramodell, einem Beleuchtungsmodell sowie einem Szenenmodell. Im weiteren Verlauf dieses Beitrags werden lediglich Objekt- und Bewegungsmodelle betrachtet.

Im Bereich der Codierung von Bewegtbildsequenzen sind unter anderem. folgende Objektmodelle geläufig:

- Blöcke mit örtlich statistisch abhängigen Bildpunkten,
- bewegte Blöcke
- bewegte unbekannte (beliebig berandete) Objekte,
- bewegte bekannte (beliebig berandete) Objekte.

Bisherige Standards [2, 8, 9, 11] beschreiben eine Bildsequenz durch Blökke mit örtlich statistisch abhängigen Bildpunkten (im Intraframe-Modus) bzw. durch starre bewegte Blöcke (im Interframe-Modus). Im Interframe-Modus wird das Bewegungsmodell zweidimensionaler translatorischer Bewegung angewendet. Daraus ist unmittelbar ersichtlich, daß bisher standardisierte Codierverfahren die natürlichen Gegebenheiten einer realen Szene äußerst unvollkommen modellieren. Somit entstehen bei der Codierung von Sequenzen mit natürlichem Szeneninhalt für einen großen Anteil der Bildfläche Prädiktionsfehler ungleich Null, die codiert und zum Empfänger übertragen werden müssen.

Bei dem in der Entwicklung befindlichen Standard MPEG-4 lassen sich durch die VOP-Struktur [19] (VOP = Video Object Plane) beliebig geformte Bildbereiche adressieren. Da gemäß der Definition im MPEG-4-Video-Verifikationsmodell [10] die VOPs Projektionen realer Objekte der realen Szene in die Bildebene darstellen sollen, läßt sich auf das MPEG-4-Video-Verifikationsmodell das Objektmodell bewegter unbekannter zweidimensionaler Objekte anwenden. Damit ist insbesondere an Objektberandungen eine genauere Beschreibung des Bildinhalts möglich.

Eine Erweiterung dieses Quellenmodells kann dadurch vorgenommen werden, daß die Objekte in der Bildebene als Projektionen dreidimensionaler Modellobjekte in die Bildebene aufgefaßt werden. Durch diese dreidimensionalen Modellobjekte sollen die an beliebigen Positionen im dreidimensionalen Raum befindlichen realen Objekte der realen Szene genauer beschrieben werden. Dabei müssen Modellobjekte und reale Objekte nicht identisch sein; es ist hinreichend, wenn die jeweiligen Projektionen in die Bildebene identisch sind. In diesem Zusammenhang wird häufig von modellbasierter Codierung gesprochen. Da dieser Begriff jedoch nicht eindeutig ist, wird er in diesem Artikel nicht weiter verwendet.

Eine weitere Steigerung der Codiereffizienz ist zu erwarten, wenn das Wissen über das Vorhandensein bestimmter Szeneninhalte ausgenutzt wird. Hierbei spricht man vom Modell bewegter bekannter Objekte. Vorgefertigte Modellobjekte werden einem Speicher entnommen und an die realen Objekte angepaßt. Die zugehörige Codiertechnik wird als wissensbasierte Codierung bezeichnet. In bestimmten Situationen, zum Beispiel bei Vorhandensein eines Gesichts, lassen sich genauere Bewegungsmodelle einsetzen, die eine noch effizientere Codierung ermöglichen. Anstatt beliebige Bewegungen zu erlauben, lassen sich in dem gegebenen Beispiel die für ein Gesicht typische Mimik durch Mimikparameter beschreiben. Ein solches Bewegungsmodell wird bei der sogenannte semantischen Codierung zugrunde gelegt.

Im Kapitel 2 dieses Artikels werden zunächst Quellenmodelle bisheriger Standards diskutiert. Das Quellenmodell, das dem MPEG-4-Video-Verifikationsmodell zugrunde liegt, wird in Kapitel 3 untersucht. Erweiterungen des MPEG-4-Quellenmodells durch die Berücksichtigung dreidimensionaler Modellobjekte und durch die Ausnutzung von Wissen über Szeneninhalte werden in den Kapiteln 4 bzw. 5 erörtert. Dabei werden mit den bekannten Codiertechniken erzielte Ergebnisse berichtet sowie die bei der semantischen Codierung zu erwartende Codiereffizienz abgeschätzt. Kapitel 6 gibt ein Beispiel für einen mehrstufigen Coder, der auf verschiedenen Quellenmodellen basierende Codiertechniken vereint und die unterschiedlichen Eigenschaften der Quellenmodelle in vorteilhafter Weise ausnutzt. Kapitel 7 schließt mit einer Zusammenfassung.

2. Quellenmodelle bisheriger Standards

Das Bildsignal digitaler Bildsequenzen wird durch Abtastung und Quantisierung des realen kontinuierlichen Kamerasignals gewonnen. Es läßt sich dann für einen bestimmten Zeitpunkt als zweidimensionale Matrix der quantisierten Abtastwerte (digitales Bild) darstellen. Dabei werden die Ortspositionen innerhalb der Matrix als Bildpunkte bezeichnet. In natürlichen Bildern herrscht zwischen benachbarten Bildpunkten eine hohe statistische Abhängigkeit. Durch eine bildpunktübergreifende Codierung kann diese Abhängigkeit zur Datenkompression ausgenutzt werden. Verschiedene Ansätze hierzu sind aus der Literatur bekannt. Außer der linearen Prädiktion mit anschließender skalarer Quantisierung für jeden einzelnen Bildpunkt basieren solche Algorithmen auf einer blockweisen Verarbeitung, zum Beispiel bei der Transformationscodierung oder der Vektorquantisierung, oder auf einer Verarbeitung in einem gegebenen Einflußbereich, zum Beispiel bei der Teilbandcodierung. In diesen Fällen können naturgemäß nur die statistischen Abhängigkeiten innerhalb eines Blockes oder innerhalb des Einflußbereichs ausgenutzt werden. Bisherige Standards [2, 8,9,11] wenden eine Codierung mit Diskreter Cosinus-Transformation (DCT) auf Blöcke einer vorgegebenen Größe von 8×8 Bildpunkten an.

In Bewegtbildsequenzen existiert nun zusätzlich eine statistische Abhängigkeit zwischen aufeinanderfolgenden Bildern, welche sich durch eine zeitliche Prädiktion zur Datenkompression ausnutzen läßt. Dazu wird das aktuelle Bild in Blöcke der Größe 16×16 oder 8×8 Bildpunkte eingeteilt und für jeden Block ***ein*** Displacementvektor geschätzt. Auf das Prädiktionsfehlersignal wird dann in der gleichen Weise eine DCT-Codierung angewendet wie im obigen Fall auf das Bildsignal. Man spricht hier vom Quellenmodell starrer translatorisch bewegter Blöcke.

Da nur ein Displacementvektor pro Block geschätzt wird, entstehen Probleme bei der Modellierung des Bildsignals, wenn in einem Block zwei unterschiedliche Bewegungen auftreten. Das geschieht grundsätzlich, wenn zwei sich unterschiedlich bewegende Objekte in einen Block projiziert werden. Da höchstens eine der zwei Bewegungen richtig wiedergegeben werden kann, können im Projektionsbereich des anderen Objekts oder beider Objekte hohe Prädiktionsfehler auftreten. Wenn insbesondere in Anwendungsfällen mit sehr niedrigen Übertragungsbitraten nicht genügend Bits zur Prädiktionsfehlercodierung zur Verfügung stehen, entstehen für diese Art der Codierung typische Codierfehler, die sogenannte Blockeffekte und Moskitoartefakte. Diese Fehler können nur durch eine genaue Wiedergabe der Berandungen sich unterschiedlich bewegender Objekte vermieden werden, wie sie durch das MPEG-4-Video-Verifikationsmodell ermöglicht wird.

3. Quellenmodell des MPEG-4-Video-Verifikationsmodells

Grundlage des MPEG-4-Video-Verifikationsmodells ist die VOP-Struktur (VOP = Video Object Plane) [19]. Eine Bildsequenz besteht dabei aus einer oder mehreren VOPs. Eine VOP kann ihrerseits ein zweidimensionales Objekt oder mehrere zweidimensionale Objekte beinhalten. **Bild 1** zeigt ein Beispiel für einen allgemeinen Szeneninhalt. Die sechs Objekte (Bildhintergrund, Haus, Baum, Auto und die zwei Personen) können einer bis sechs VOPs zugeordnet werden. Es ist möglich, mehrere dieser Objekte in einer VOP zusammenzufassen. Gemäß der Definition im Verifikationsmodell sollen die zweidimensionalen Objekte Projektionen realer dreidimensionaler Objekte der realen Szene in die Bildebene darstellen [10]. Es läßt sich daher auf das MPEG-4-Video-VM das Objektmodell bewegter zweidimensionaler Objekte anwenden. Damit können inhaltsabhängige Funktionalitäten realisiert werden, wenn als Inhalte eben diese zweidimensionalen Objekte angesehen werden. Da die zweidimensionalen Objekte beliebig geformt sein dürfen, können diese Funktionalitäten bildpunktgenau angewendet werden. Außerdem lassen sich die Silhouetten der realen Objekte und auch die Bewegungen insbesondere dort, wo nach der Projektion in die Bildebene zwei unterschiedlich bewegte zweidimensionale Objekte aneinanderstoßen, genau wiedergeben.

Es wurden für das Objektmodell bewegter zweidimensionaler Objekte zwei Bewegungsmodelle untersucht [7]: das

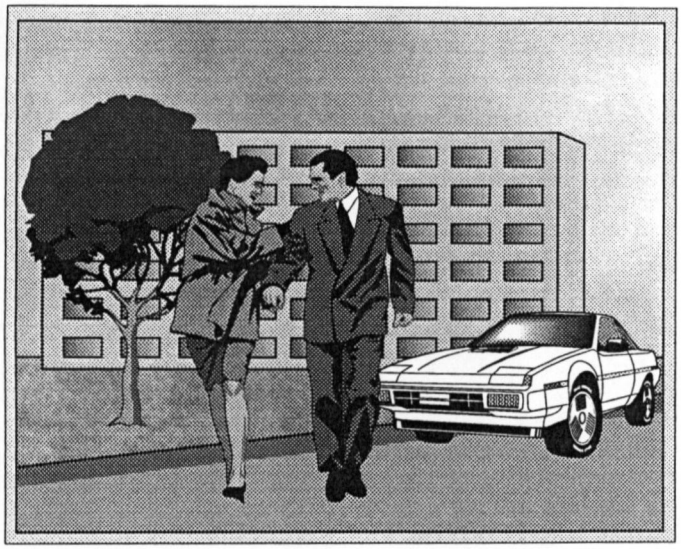

Bild 1. Beispiel für einen natürlichen Bildinhalt zur Illustration der VOP-Generierung. Die sechs Objekte (einschließlich Bildhintergrund) können einer bis sechs VOPs zugeordnet werden. Mehrere Objekte können in einer VOP zusammengefaßt werden

Modell dreidimensionaler affiner Bewegung (dabei werden die Objekte als starr angenommen), und das Modell zweidimensionaler translatorischer Bewegung (bei dem die Objekte als flexibel angenommen werden). Starr bedeutet, daß die Bewegung eines Objekts mit *einem* einzigen Bewegungsvektor beschrieben wird. Flexibel bedeutet, daß die Bewegung eines Objekts mit *mehreren* Bewegungsvektoren beschrieben wird.

Bei dem Modell dreidimensionaler affiner Bewegung starrer Objekte werden die Objekte als unendlich dünne Ebenen, die sich beliebig im dreidimensionalen Raum bewegen, betrachtet. Die Bewegung jedes Objekts wird beschrieben durch einen Bewegungsvektor bestehend aus acht Abbildungsparametern, die mittels einer Regressionsanalyse berechnet werden können. Eine Reduktion des Parametersatzes auf zum Beispiel vier oder zwei Abbildungsparameter pro Objekt ist möglich. Dabei wird ebenfalls die Menge der beschreibbaren Bewegungen reduziert. Zum Beispiel können im Falle der Verwendung von zwei Abbildungsparametern nur zweidimensionale translatorische Bewegungen des Objekts modelliert werden. Komplexere Bewegungen müssen dann durch zweidimensionale translatorische Bewegungen approximiert werden.

Wird das Modell zweidimensionaler translatorischer Bewegung flexibler Objekte angewendet, läßt sich die Bewegung durch ein Displacementvektorfeld beschreiben, das jedem Bildpunkt genau einen Displacementvektor zuordnet. Dazu wird eine Displacementschätzung eingesetzt, die die von der Bewegung lokaler Bereiche des realen Objekts verursachten Verschiebungen von Teilen des Bildes schätzt. Um die Datenrate für die Displacementvektoren gering zu halten, wird nur jeder N-te Vektor in jeder N-ten Zeile codiert und übertragen. Die fehlenden Vektoren lassen sich beim Empfänger dann so ermitteln, daß allen Bildpunkten in einem N×N-Block derselbe Vektor zugeordnet wird. Alternativ läßt sich hier in vorteilhafter Weise auch eine Bilinearinterpolation einsetzen [1,6].

Beide Bewegungsmodelle wurden in [7] vergleichend bewertet. Das Modell dreidimensionaler affiner Bewegung angewendet auf starre Objekte beschreibt eine einzelne Bewegung eines Objekts genauer als das Modell zweidimensionaler translatorischer Bewegung angewendet auf flexible Objekte, andererseits können lokal unterschiedliche Bewegungen in einem Objekt nicht modelliert werden. So steht bei jedem Modell ein Vorteil einem Nachteil gegenüber. Die Untersuchungen ergaben, daß zumindest in typischen Bildtelefonsequenzen der Effekt der lokal unterschiedlichen Bewegungen in einem Objekt überwiegt, so daß das Modell translatorischer Bewegung flexibler Objekte in diesem Fall zu einer geringeren Prädiktionsfehlerleistung führt. Bei der Realisierung eines objektbasierten Coders für sehr niedrige Übertragungsbitraten nach [6] wurde daher dieses Modell eingesetzt. Ebenso gestattet das MPEG-4-Video-VM die Berücksichtigung der Flexibilität von Objekten.

Alternativ zu dem beschriebenen Quellenmodell bewegter Objekte können bei dem MPEG-4-Video-VM auch andere Quellenmodelle angewendet werden, zum Beispiel das Modell von Regionen ähnlicher Textur [13]. Man spricht in diesem Zusammenhang von regionenbasierter Codierung. Dabei gibt die Unterteilung der Bildebene im allgemeinen nicht die Projektion der realen Objekte wieder. Auf diese Modelle wird im weiteren Verlauf dieses Artikels nicht weiter eingegangen.

Unabhängig vom gewählten Quellenmodell kann selbstverständlich jedes 2D-Objekt oder jede Region weiter in Unterregionen aufgeteilt werden, wenn sich die Codierung der zusätzlichen Forminformation im Hinblick auf eine höhere Codiereffizienz als lohnenswert zeigt. Zwar ist diese Art der Codierung in der Syntax der augenblicklichen Version des MPEG-4-Video-VMs nicht vorgesehen. Sie wird aber in späteren Versionen mit Gewißheit berücksichtigt werden, wenn sie sich im Rahmen der Core-Experimente als vorteilhaft erweist.

4. Quellenmodelle mit dreidimensionalen unbekannten Objekten

Eine Erweiterung des MPEG-4-Video-Verifikationsmodells stellt die Berücksichtigung der dritten Dimension bei der Objektmodellierung dar. Ein Vorteil des Modells dreidimensionaler Objekte gegenüber dem Modell zweidimensionaler Objekte ist die der Natur entsprechend genauere Modellierung der realen Objekte, wodurch insgesamt eine genauere Prädiktion erwartet werden kann. Außerdem hat das Modell dreidimensionaler Objekte den Vorteil, daß bei Drehungen des Objekts Teile, die zuvor auf der Hinterseite lagen und durch die Drehung sichtbar werden, besser prädiziert werden können, soweit sie im bisherigen Verlauf der Szene schon einmal sichtbar waren. Dazu wird die Textur dieser Teile, wenn sie erstmalig sichtbar sind, den entsprechenden Objektbereichen zugeordnet, sog. "Texture Mapping". Sie steht dann bei jedem weiteren Sichtbarwerden im Verlaufe der Szene zur Prädiktion zur Verfügung. Da eine Kombination des Objektmodells dreidimensionaler Objekte mit einem Bewegungsmodell zweidimensionaler Bewegung nicht sinnvoll erscheint, wird es grundsätzlich mit einem Bewegungsmodell dreidimensionaler Bewegung kombiniert.

In [18] wird das Quellenmodell starrer dreidimensionaler Objekte mit dreidimensionaler Bewegung untersucht und für Bildtelefonsequenzen in CIF-Auflö-

sung mit dem Modell flexibler zweidimensionaler Objekte mit zweidimensionaler translatorischer Bewegung verglichen. Da bei der dortigen Realisierung kein Wissen über den Szeneninhalt ausgenutzt wird, wird die Form eines dreidimensionalen Modellobjekts geschätzt und durch ein Dreiecksgitter dargestellt. Es zeigt sich, daß ab einer hinreichend hohen Übertragungsbitrate, die bei Bildtelefonsequenzen in CIF-Auflösung um 64 kbit/s liegt, das Quellenmodell starrer dreidimensionaler Objekte mit dreidimensionaler Bewegung dem flexibler zweidimensionaler Objekte mit zweidimensionaler translatorischer Bewegung überlegen ist. Bei niedrigeren Übertragungsbitraten führt allerdings letztgenanntes zu einer höheren Codiereffizienz.

Da die Annahme starrer Objekte die natürlichen Gegebenheiten nur stark vereinfacht wiedergibt, sind zwei Ansätze zur genaueren Modellierung bekannt. In [18] wird ein Quellenmodell flexibler dreidimensionaler Objekte mit dreidimensionaler Bewegung vorgeschlagen. Um den Anstieg der Bitmenge für die Bewegungsinfomation möglichst gering zu halten, werden lokale Bewegungsvektoren im Gegensatz zum Modell flexibler zweidimensionaler Objekte mit zweidimensionaler translatorischer Bewegung nicht im gesamten Objekt geschätzt, sondern nur dort, wo die globale Bewegungsbeschreibung nicht ausreichend ist. Außerdem wird in der genannten Realisierung für die lokale Bewegung keine allgemeine dreidimensionale Bewegung geschätzt, sondern eine zweidimensionale Bewegung tangential zur Objektoberfläche. Wegen der erhöhten Bitmenge für die Bewegungsinfomation zeigt sich dieser Ansatz erst bei relativ hohen Übertragungsbitraten als vorteilhaft, bei den untersuchten Bildtelefonsequenzen in CIF-Auflösung ab etwa 54 kbit/s. Bei geringeren Übertragungsbitraten stehen zur Prädiktionsfehlercodierung überproportional weniger Bits zur Verfügung, so daß die Codiereffizienz gegenüber dem Modell starrer Objekte überproportional stark abnimmt.

In [15] wird ein Quellenmodell gegliederter dreidimensionaler Objekte mit dreidimensionaler Bewegung vorgeschlagen. Dieser Ansatz geht von der Annahme aus, daß insbesondere in Kopf-Schulter-Szenen sich die Objekte zwar nicht als Ganzes starr bewegen, aber doch Teile der Objekte. Ein Objekt besteht hierbei aus starren Komponenten, die ihrerseits flexibel miteinander verbunden sind. Außerdem gestattet dieser Ansatz die Modellierung sich verdeckender Komponenten. Da bei diesem Ansatz für jede Komponente sowohl die Form als auch ein dreidimensionaler Bewegungsvektor geschätzt werden, erhöht sich die Bitmenge für Form und Bewegung. Demgegenüber steht eine stark verbesserte Prädiktion der Farbinformation. Erste Ergebnisse zeigen für typische Bildtelefonsequenzen im Bereich von 50 bis 60 kbit/s einen Gewinn von etwa 15% gegenüber dem Modell starrer Objekte. Erweiterungen dieser Codiertechnik sowie eine vollständige Bewertung befinden sich noch in der Bearbeitung.

5. Quellenmodelle mit dreidimensionalen bekannten Objekten

Ist das Vorhandensein eines bestimmten realen Objekts in der Szene bekannt, kann dieses Wissen zur genaueren Modellierung ausgenutzt werden. Entsprechende Codiertechniken werden wissensbasierte Codierung genannt. zum Beispiel kann bei Bildtelefonanwendungen im allgemeinen davon ausgegangen werden, daß mindestens ein Gesicht in der Szene vorhanden ist. Ist das der Fall, kann ein in einem Speicher abgelegtes Modellobjekt, hier eine vorher gefertigte Gesichtsmaske, zum Beispiel [20], an das reale Gesicht angepaßt werden [12]. Hierzu muß eine Anpassung der Gesichtsmaske an die Position und die Größe des realen Gesichts erfolgen. Durch ein sogenanntes „Face Tracking" [22] wird die Position und Größe der Gesichtsmaske Bild für Bild nachgeführt und ständig neu adaptiert. Ist die Position des Gesichts hiernach bekannt, kann zwischen Gesichtsbereich und anderen Bildbereichen unterschieden werden. Dadurch ist eine subjektiv gewichtete Bitallokation möglich. Durch diese Bitallokation und durch die genauere Modellierung des realen Gesichts mittels der Gesichtsmaske wird eine Datenrateneinsparung von 15 bis 20% erzielt.

Ist nun ein Objekt in der Szene gefunden worden, das durch ein gespeichertes Modellobjekt modelliert werden kann, ist es unter Umständen möglich, die für das Objekt typische Bewegung zu modellieren, anstatt allgemein Displacements oder dreidimensionale Bewegungen zu schätzen. Im Falle der Modellierung eines Gesichts werden die Bewegungen in Form von Mimikparametern geschätzt. Da die Bewegung in diesem Fall in semantischem Zusammenhang mit dem Objekt steht, kann eine Codierung, die eine solche Modellierung durchführt, als semantische Codierung bezeichnet werden [3,14]. Ein vollständiger semantischer Coder ist bislang noch nicht vorgestellt worden. In [5] werden allerdings Bitmengen angegeben, die zur Codierung der Information im Gesichtsbereich erforderlich sind. Die insgesamt zu erwartende Codiereffizienz läßt sich abschätzen, indem außerhalb des Gesichtsbereichs Bitmengen aus bekannten Codiertechniken abgeleitet werden [16]. Diese Abschätzung führt auf eine Datenrateneinsparung um den Faktor 2 bis 3 gegenüber Codiertechniken ohne semantische Codierung. Damit wäre die Codierung von Bildtelefonsequenzen in CIF-Auflösung mit akzeptabler Bildqualität bei etwa 20 kbit/s, in QCIF-Auflösung bei etwa 5 kbit/s möglich. **Tabelle I** gibt eine Übersicht für mehrere Quellenmodelle über die Aufteilung der Datenraten auf Form-, Bewegungs- und Farbinformation gemittelt über mehrere Bildtelefonsequenzen in CIF-Auflösung.

6. Beispiel für einen mehrstufigen Coder

In [16] wird ein mehrstufiger Coder vorgestellt, der durch eine Zusammenschaltung die Vorteile verschiedener Codiertechniken vereint. **Bild 2** zeigt das Blockschaltbild der Anordnung. In der Stufe I wird durch eine Transformationscodierung das Quellenmodell von Blöcken mit örtlich statistisch abhängigen Bildpunkten angewendet. Durch eine zusätzliche zeitliche Prädiktion wird das Quellenmodell bewegter Blökke in der Stufe II mittels einer Hybrid-Codierung realisiert. Dadurch daß hier ein weiterer Informationsstrang erzeugt wird, ist die Verwendung dieser Stufe nur lohnenswert, wenn die zusätzliche Bitmenge durch die Einsparung bei der Prädiktionsfehlercodierung der Farbinformation mindestens kompensiert wird; anderenfalls muß zur Stufe I zurückgeschaltet werden.

In heutigen Standards [2,8,9,11] sind diese beiden Stufen enthalten. Beim Hinzufügen einer objektbasierten Codierung als

Tabelle I. Aufteilung der Gesamtdatenrate auf die drei Informationsstränge Bewegung, Form und Farbe bei gleicher Bildqualität für fünf auf unterschiedlichen Quellenmodellen basierende Codiertechniken nach [16] (Mittelwerte in bit/frame über typische Bildtelefonsequenzen, Bildauflösung: CIF, Datenrate für die Prädiktionsfehler der Farbinformation zu 1,1 bit/pel angesetzt), für semantische Codierung geschätzt

Codiertechnik	Bewegung	Form	Farbe	Gesamt-Datenrate
ObjektbasierteCodierung (flexible 2D-Objekte)	1100	900	4400	6400
ObjektbasierteCodierung (starre 3D-Objekte)	200	1700	4070	5970
ObjektbasierteCodierung (flexible 3D-Objekte)	650	1550	3080	5280
Wissensbasierte Codierung	350	1550	2750	4650
Semantische Codierung	290	850	900	2040

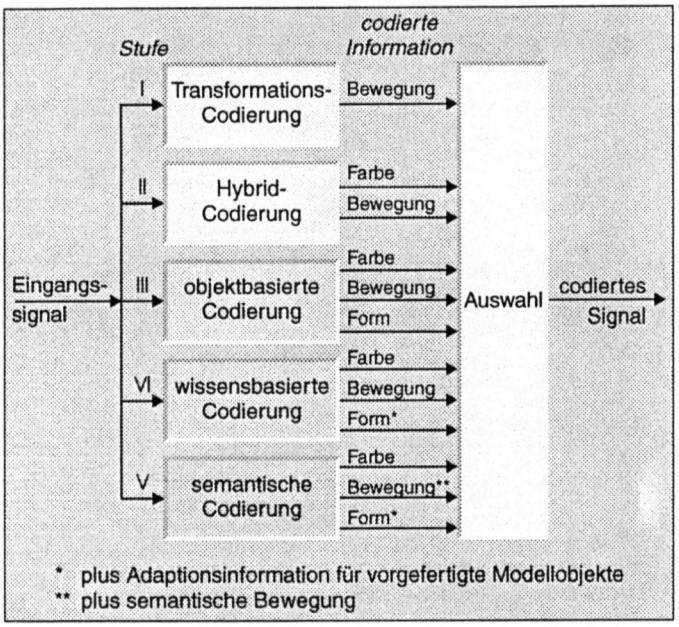

Bild 2. Blockschaltbild einer mehrstufigen Codierung nach [16]

Stufe III wird ebenfalls ein neuer Informationsstrang, nämlich die Forminformation, erzeugt. Die Verwendung dieser Stufe ist daher nur sinnvoll, wenn die hierfür erforderliche Bitmenge durch eine Einsparung in den Bitmengen für die Bewegungs- und Farbinformation, insbesondere durch die genauere Prädiktion an Objekträndern, mindestens kompensiert wird, es sei denn, daß die bildpunktgenaue Adressierung von Bildinhalten für inhaltsabhängige Funktionalitäten gegenüber der Codiereffizienz im Vordergrund steht.

Ist das Vorhandensein bestimmter Szeneninhalte bekannt, kann dieses Wissen durch eine wissensbasierte Codierung als Stufe IV ausgenutzt werden. Ist kein so modellierbares Objekt vorhanden oder wird die zusätzliche Bitmenge für die Adaption des vorgefertigten Modellobjekts nicht durch eine Einsparung in den anderen Informationssträngen kompensiert, muß zu einer der Stufen I-III zurückgeschaltet werden. Kann die für ein Modellobjekt typische Bewegung im Rahmen einer semantischen Codierung erfolgreich modelliert werden, kann diese als Stufe V realisiert werden. Ist die Modellierung nicht erfolgreich, das heißt der im Bereich des Modellobjekts verbleibende Prädiktionsfehler der Farbinformation ist höher als bei einer Schätzung von Displacements oder dreidimensionalen Bewegungen, die nicht im semantischen Zusammenhang mit dem Objekt stehen, muß wiederum zu einer der anderen Stufen zurückgeschaltet werden.

Ein besonderes Problem bei einem mehrstufigen Coder ist die Stufenauswahl. Die Auswahl zwischen den Stufen I und II ist implizit in der Realisierung heutiger Standards enthalten. Für die Auswahl zwischen den Stufen II und III wurde ein erster einfacher Ansatz in [4] vorgestellt. In [21] wird ein Ansatz vorgeschlagen, der von einer Rate-Distortion-Analyse hergeleitet wurde. Kriterien für die Auswahl zwischen den Stufen III-V sind noch in der Entwicklung.

Allgemein kann gesagt werden, daß von Stufe I bis Stufe V die Quellenmodelle ansteigend spezialisiert sind, während in umgekehrter Reihenfolge die Quellenmodelle allgemeiner einsetzbar sind. Das bedeutet, daß bei Vorliegen einer Situation, die von einem komplexeren Quellenmodell modelliert werden kann, dieses zu einer effizienteren Codierung führt als Quellenmodelle unterer Stufen. Ist diese Voraussetzung nicht erfüllt, führen Quellenmodelle unterer Stufen zu einer effizienteren Codierung, da komplexere Quellenmodelle mehr Informationsstränge erzeugen, ohne daß eine entsprechende Kompensation in anderen Informationssträngen auftritt.

7. Zusammenfassung

Das Quellenmodell, das dem MPEG-4-Video-Verifikationsmodell zugrunde liegt, wurde diskutiert. Es gestattet die Modellierung zweidimensionaler flexibler Objekte, die beliebig berandet sein dürfen. Gegenüber bisherigen Standards, die eine Verarbeitung von Blöcken in einem vorgegebenen Raster durchführen, ist hiermit eine genauere Beschreibung der Objektformen und der Bewegungen, insbesondere an Objekträndern, und dadurch eine genauere Prädiktion des Bildsignals möglich. Außerdem gestattet es die bildpunktgenaue Adressierung von Bildinhalten, wodurch inhaltsabhängige Funktionalitäten bildpunktgenau realisiert werden können.

Erweiterungen dieses Quellenmodells, die bei in Entwicklung befindlichen Codiertechniken angewendet werden, wurden vorgestellt. Erzielte Datenraten wurden angegeben für die Quellenmodelle dreidimensionaler unbekannter sowie dreidimensionaler bekannter Objekte, die in der 3D-objektbasierten bzw. in der wissensbasierten Codierung angewendet werden. Die 3D-objektbasierte Codierung starrer oder flexibler Objekte ergibt gegenüber der 2D-objektbasierten Codierung, die im MPEG-4-Video-VM durchgeführt werden kann, einen Vorteil in der Codiereffizienz, wenn eine hinreichend hohe Übertragungsbitrate zur

Verfügung steht. Bei niedrigeren Übertragungsbitraten zeigt sich die 2D-objektbasierte Codierung vorteilhaft, es sei denn, daß die genauere Modellierung der realen Objekte ausgenutzt werden soll, zum Beispiel in der Videosynthese von gemischt natürlichen und synthetischen Bildinhalten im Rahmen von SNHC (Synthetic Natural Hybrid Coding). Für eine Codierung basierend auf dem Quellenmodell gegliederter dreidimensionaler Objekte liegen erste Ergebnisse vor, die für typische Bildtelefonsequenzen im Bereich um 50 bis 60 kbit/s eine Datenratenreduktion von etwa 15 % ergeben. Die wissensbasierte Codierung erzielt für eine bestimmte Realisierung [12] bei Bildtelefonsequenzen eine Datenrateneinsparung von 15 bis 20% bei gleicher Bildqualität.

Für eine semantische Codierung wurde eine Abschätzung der Datenrate gemäß [16] angeführt. Ausgehend von bekannten Teilergebnissen kann eine Datenratenreduktion um den Faktor 2 bis 3 erwartet werden gegenüber der Codierung ohne semantisches Bewegungsmodell. Damit wäre zum Beispiel die Codierung von Bildtelefonsequenzen in akzeptabler Bildqualität mit etwa 20 kbit/s bei CIF-Auflösung bzw. 5 kbit/s bei QCIF-Auflösung möglich. Die Verifikation dieser Abschätzung befindet sich zur Zeit in der Untersuchung.

Ein Beispiel für einen mehrstufigen Coder wurde zitiert [16], bei dem mehrere Codiertechniken, die auf unterschiedlichen Quellenmodellen basieren, verknüpft sind. Dabei wurde die Problematik der Auswahl der effizientesten Codiertechnik besonders hervorgehoben. Diese Problematik ist zwar nicht für die Standardisierung, die nur Bitstromsyntax und Decoder festlegt, von Bedeutung, wohl aber für die Effizienz des Coders. Eine Lösung für die Auswahl zwischen blockbasierter und objektbasierter Codierung auf der Basis einer Rate-Distortion-Analyse wurde bereits in der Literatur vorgestellt [21]. Eine Verfeinerung dieser Lösung sowie Kriterien für die Auswahl zwischen objektbasierter, wissensbasierter und semantischer Codierung sind zur Zeit in Untersuchung.

Der Autor möchte sich an dieser Stelle bei Prof. Dr.-Ing. H.G. Musmann sowie bei den Kollegen Dipl.-Ing. M. Kampmann und Dipl.-Ing. M. Wollborn für die hilfreichen Diskussionen bedanken, die zum Entstehen dieses Artikels beigetragen haben.

Schrifttum

[1] Buschmann, R.: Efficiency of diplacement estimation techniques", eingesendet zur Veröffentlichung in "Signal Processing: Image Communication"

[2] CCITT Recommendation H.261, Video codec for audiovisual services at p * 64 kbit/s, 1989.

[3] Choi, C.S., Aizawa, K.; Harashima, H.; Takebo, T.: Analysis and synthesis of facial image sequences in model-based image coding. IEEE Transact. on Circ. and Syst. in Video Techn.: Special Issue on Very Low Bit Rate Video Coding, Vol. 4 (1994), No. 3, pp. 257-275.

[4] Chowdhury, M..FClark, A.F.; Downton, A.C.; Morimatsu, E.; Pearson, D.C.: A switched model-based coder for video signals. IEEE Trans. on Circ.and Syst. for Video Techn: Special Issue on Very Low Bit Rate Video Coding, Vol. 4 (1994), No. 3, pp. 216-227.

[5] Fischl, J.; Miller, B.; Robinson, J.: Parameter tracking in a muscle-based analysis-synthesis coding system. Proc. Picture Coding Symposium '93, Lausanne, Switzerland, pp. 2.3.1-2.3.2.

[6] Gerken, P.: Object-Based Analysis-Synthesis Coding of Image Sequences at Very Low Bit Rates. IEEE Transact. on Circ. and Syst. in Video Techn.: Special Issue on Very Low Bit Rate Video Coding, Vol. 4 (1994), No. 3, pp. 228-235.

[7] Hötter, M.: Object-oriented analysis-synthesis coding based on moving two-dimensional objects. Signal Processing: Image Communication, Vol. 2 (1990), No. 4, pp. 409-428.

[8] ISO/IEC IS 11172: Information technology Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s. nternational Standard, 1993.

[9] ISO/IEC IS 13818 I ITU-T Recommendation H.262: Information technology — Generic coding of moving pictures and associated audio information. International Standard, 1995.

[10] ISO/IEC, MPEG-4 video verification model version 2.0. Doc. ISO/IEC JTC1/SC29/WG11 N1260, March 1996.

[11] ITU-T Recommendation H.263: Video coding for low bit rate communication, 1995.

[12] Kampmann, M.; Ostermann, J.: Automatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer. angenommen zur Veröffentlichung in Signal Processing: Image Communication

[13] Kunt, M.; Ikonomopolous, A.; Kocher, M.: Second-generation image coding techniques. Proc. IEEE, Vol. 73 (1985), No. 4, pp. 549-574.

[14] Li, H.; Roivainen, P.; Forchheimer, R.: 3-D motion estimation in model-based facial image coding. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15 (1993), No. 6, pp. 545-555.

[15] Martínez, G.: Shape estimation of moving articulated 3-D objects for object-based analysis-synthesis coding (OBASC)", angenommen zur Veröffentlichung in Signal Processing: Image Communication

[16] Musmann, H.G.: A layered coding system for very low bit rate video coding. Signal Processing: Image Communication, Vol. 7 (1995), Nos. 4-6, pp. 267-278.

[17] Musmann, H.G.; Hötter, M.; Ostermann, J.: Object-oriented analysis-synthesis coding of moving images, Signal Processing: Image Communication, Vol. 1 (1989), No. 2, pp. 117-138.

[18] Ostermann, J.: Analyse-Synthese-Codierung basierend auf dem Modell bewegter dreidimensionaler Objekte. Dissertation, Universität Hannover, 1995.

[19] Pereira, F.: MPEG-4: a new challenge for the representation of audio-visual information" (keynote speech). Proc. Picture Coding Symposium '96, Melbourne, Australia, pp. 7-16.

[20] Rydfalk, M.: CANDIDE: a parameterised face. Internal Report, University of Linköping, Sweden, 1987.

[21] Wollborn, M.: Investigations on an object-based layered coding system. Proc. Picture Coding Symposium '96, Melbourne, Australia, pp. 531-533.

[22] Zhang, L.: Tracking a face for knowledge-based coding of videophone sequences, eingesendet zur Veröffentlichung in Signal Processing: Image Communication.