Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

Forschungsvorhaben:
Entwicklung eines „Motion-CELP-Systems“ zur Video-Codierung inklusive Performance-Vergleich gegenüber anderen Standards
als Magister-Arbeit

Gliederung des Informationsblattes:

1.] Motivation:

In heutigen Standards zur Bewegtbildcodierung wird zumeist auf Transformationscodierung nach erfolgter Bewegungskompensation zurückgegriffen. Im bekannten H.264-Standard der JVG geschieht dies bspw. mittels Integer-DCT in 4x4 – Blöcken.
Da die Transformationscodierung, genauso wie die prädiktive Codierung, dadurch wirkt, dass sie Korrelationen beseitigt, ist ihre Anwendung auf Prädiktionsfehlerbilder, die nach einer solch umfangreichen und ausgeklügelten Bewegungskompensation, wie sie H.264 nutzt, stark dekorreliert sind, zumindest hinterfragungswürdig. Auch trifft die Annahme, dass die Bildsignale überwiegend tieffrequente Anteile in sich tragen, für dekorrelierte Bilder nicht mehr zu. In [Shu] wurde daher vorgeschlagen, in einem solchen Fall gänzlich auf die Transformation zu verzichten [Strutz].
War die Bewegungskompensation, z.B. aufgrund nicht vorhersagbarer Bildinhalte, wenig erfolgreich, so verbleiben Rest-Korrelationen, weshalb die Transformationscodierung dennoch angewendet wird.

Ausgehend eines vertieften Studiums der Sprachcodierung, fiel auf, dass es keinen ausschließenden Grund dafür gibt, das innerhalb der Sprachcodierung bahnbrechend erfolgreiche CELP-Verfahren nicht auch auf mehrdimensionale Signale anzuwenden.
Recherchen ergaben, dass Eric Dubois (Universität Quebec) 1990 und 1996 Papers über ein 2-D-CELP-System zur Codierung von Einzelbildern veröffentlichte, wobei sich in [Dubois2] eine Performance-Verbesserung gegenüber JPEG (DCT-basiert) im untersuchten Bitraten-Bereich ergab.

Wie auch immer, mag es Gründe dafür geben, dass dieser Forschungszweig nicht weiter verfolgt wurde, doch vermögen diese evtl. darin begründet zu sein, dass in weit größerer Fülle im Bereich der TC-basierten Verfahren geforscht wurde und diese, z.B. durch Entwicklung von anschließender CAVLC und CABAC, sehr effizient wurde.

Das angestrebte Forschungsvorhaben soll nun das CELP-Prinzip in die Bewegtbildcodierung einbinden und ergründen, inwieweit ein solches Motion-CELP-System gegenüber heutigen Standards bestehen kann.

Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

2.] Inhalt der Arbeit:

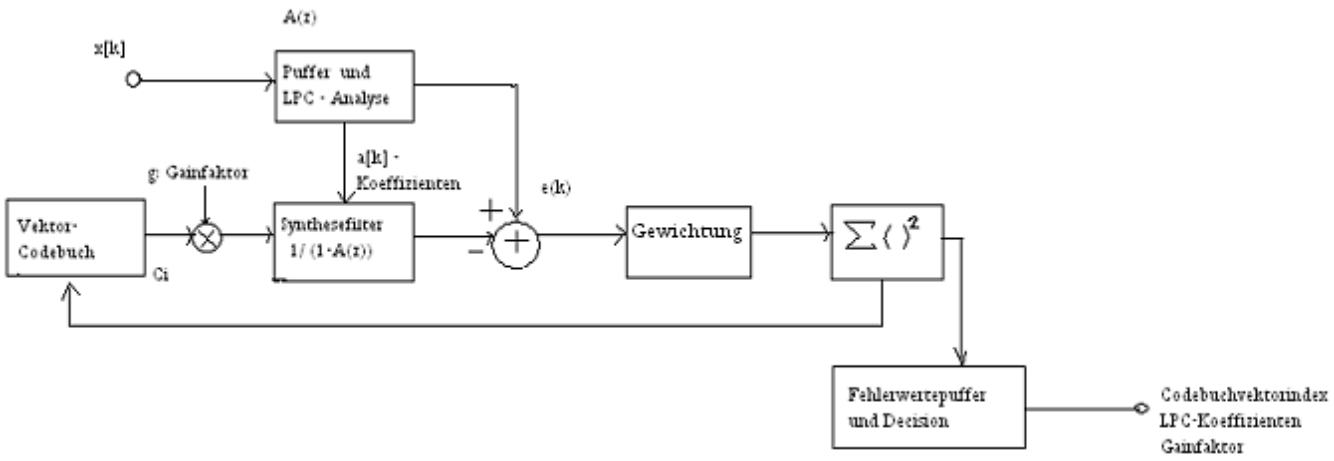
- Einarbeitung in die Bewegtbildcodierung, insbes. H.264 als Referenz-Standard
- Einarbeitung in das 2-D-CELP-System für Einzelbildcodierung
- Entwicklung und Optimierung des Motion-CELP-Systems
- Beurteilung des entwickelten Systems bezgl. R(D) und Vergleich mit anderen Standards
- Erörterung möglicher Anwendungsgebiete
- Dokumentation der Arbeit, sowie Implementierung des Systems in MatLab und/oder C/C++

Nicht-Inhalt der Arbeit:

- Es ist nicht erforderlich, dass das entwickelte System hinsichtlich irgendeines Gesichtspunktes den H.264 Standard übertrumpft.
- Es wird keine Analyse zur Wirtschaftlichkeit angestellt
- Es wird keine Optimierung bezüglich der Laufzeit angestellt. Eventuelle Fast-Search-Algorithmen für die Codebuch-Suche und/oder Bewegungsschätzung werden genannt und erklärt, falls sie Verwendung finden, jedoch werden diese der Literatur entnommen und nicht weitergehend optimiert.

3.] Das CELP-Sprachcodierungsverfahren [Kurzfassung]

Ebenso wie ADPCM nutzt das CELP-Verfahren LPC, um das Original zu dekorrelieren. Prinzipiell kann es auch aus dem ADPCM-Verfahren abgeleitet werden. Mittels eines linearen Filters wird ein Schätzwert für den aktuell anliegenden Eingangswert aus vergangenen Werten gebildet und vom Original subtrahiert. Das resultierende Differenzsignal ist nahezu weiß und weist im Mittel sehr viel geringere Varianz auf, als das Original, worin der Codierungsvorteil liegt.
Im ADPCM-Verfahren wird dieses Fehlersignal innerhalb der Filterschleife quantisert, wodurch eine Fehlerrückkopplung stattfindet, die das System bei sehr niedrigen Bitraten ineffizient macht. Es gelingt dann nicht mehr ein optimales (möglichst dekorreliertes) Differenzsignal zu erzeugen.
Hier setzt CELP an, indem es geeignete Anregungsvektoren in einem Codebuch ablegt und derjenige unter ihnen gewählt wird, der innerhalb einer im Encoder implementierten lokalen Dekodierung den kleinsten (frequenzgewichteten) quadratischen Fehler liefert.
Der Index des Codebuch-Vektors, die LPC-Koeffizienten des linearen Prädiktors, sowie ggf der Gainfaktor, mit welchem der Eintrag multipliziert wird, werden dem Empfänger übermittelt.



Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

4.] Das 2-D-CELP-System nach E. Dubois

Das Prinzip ist dem eindimensionalen Fall gleich, nur dass nicht für jeden Block die optimalen Koeffizienten des Prewhitening-Filters übertragen werden, sondern der Index des besten von K [K = 5 - 8] festen spatialen Prädiktoren.
Der Eingangsblock wird zunächst mit jedem der Prewhitening-Filter prädiziert. Mit MSE-Kriterium wird dabei der beste Prädiktor ausgewählt, dessen inverses Filter $H(z) = 1 / (1 – P(z))$ dann als Synthesefilter innerhalb der lokalen Dekodierung dient. Prinzipiell wird dieses sodann mit jedem einzelnen Codebuchvektor angeregt und derjenige unter ihnen wird gewählt, der den kleinsten Fehler liefert. Der resultierende Fehler wird als Quantisierungsfehler hingenommen.

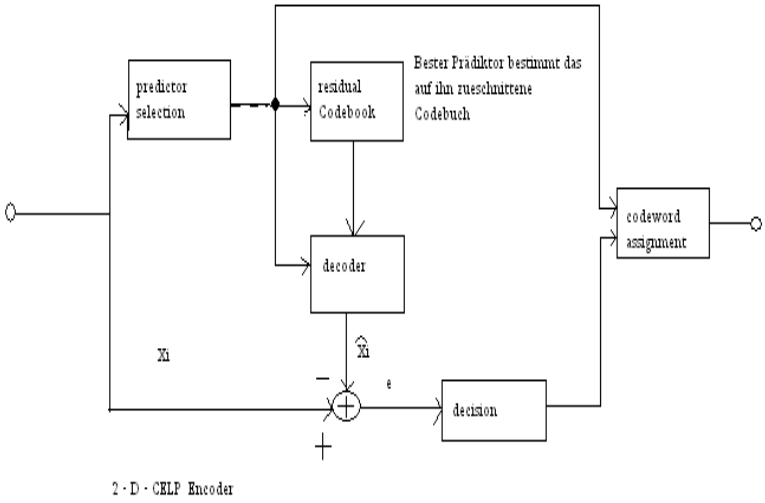
Die Prädiktoren: Da es sich bei dem Verfahren um Einzelbildcodierung handelt, sind alle Prädiktoren spatial und, gemäß CELP, linear. Ausgehend von initialen Prädiktoren, die jeweils unterschiedliche Orientationen aufweisen, wird, ähnlich dem LBG-Algorithmus zur Codebuchgenerierung, mit Trainingsvektoren geclustert und die Koeffizienten so nachgeregelt, dass der Prediction-Gain innerhalb eines jeden Clusters maximal und der MSE minimal wird.
Bei einem derart beschränkten Set von maximal 8 Prädiktoren resultierten aus dieser Art der Mittelung sehr einfache Filterstrukturen, bei welchen nur nächste Nachbarn und Linearkombinationen dieser zur Prädiktion verwendet werden, was auch zu erwarten war. Die Filterordnung war in jedem Fall 4.

Da die meisten natürlichen Bilder sowohl detailreiche als auch wenig detaillierte Gebiete aufweisen, wurde in [Dubois2] ein blockadaptives Verfahren implementiert. Über einen gegebenen Threshold wird entschieden, ob der jeweilige Block in Subblöcke aufgeteilt wird, wodurch sich die Möglichkeit der Prädiktion mittels einem der festen Prädiktoren gesteigert wird. Bei Verwendung von Codebüchern gleichen Umfangs, wird dabei zusätzlich die Genauigkeit der Vektorquantisierung heraufgesetzt, während die Verwendung größerer Blöcke hinsichtlich der Bitrate effizienter ist.

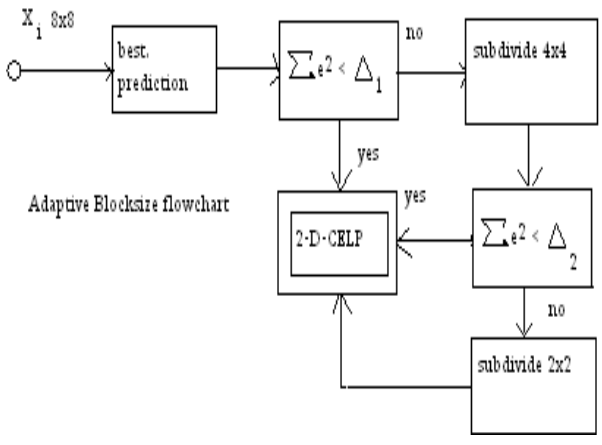
Die Codebücher werden für jeden Blockgrößenfall und ohne weiteren Overhead für jeden spatialen Fall durch sukzessives Clustering von Trainingsvektoren erstellt.

Der Blockgrößenindex, Prädiktorindex und Codebuchvektorindex werden für jeden Block dem Empfänger mittels Huffmancoding übermittelt.

In den Versuchsreihen wurde das DCT-basierte JPEG-Verfahren bei niedrigen Bitraten [0,27-0,7 bpp] im PSNR-Vergleich um durchschnittlich 1,5 dB geschlagen.



2 - D - CELP Encoder



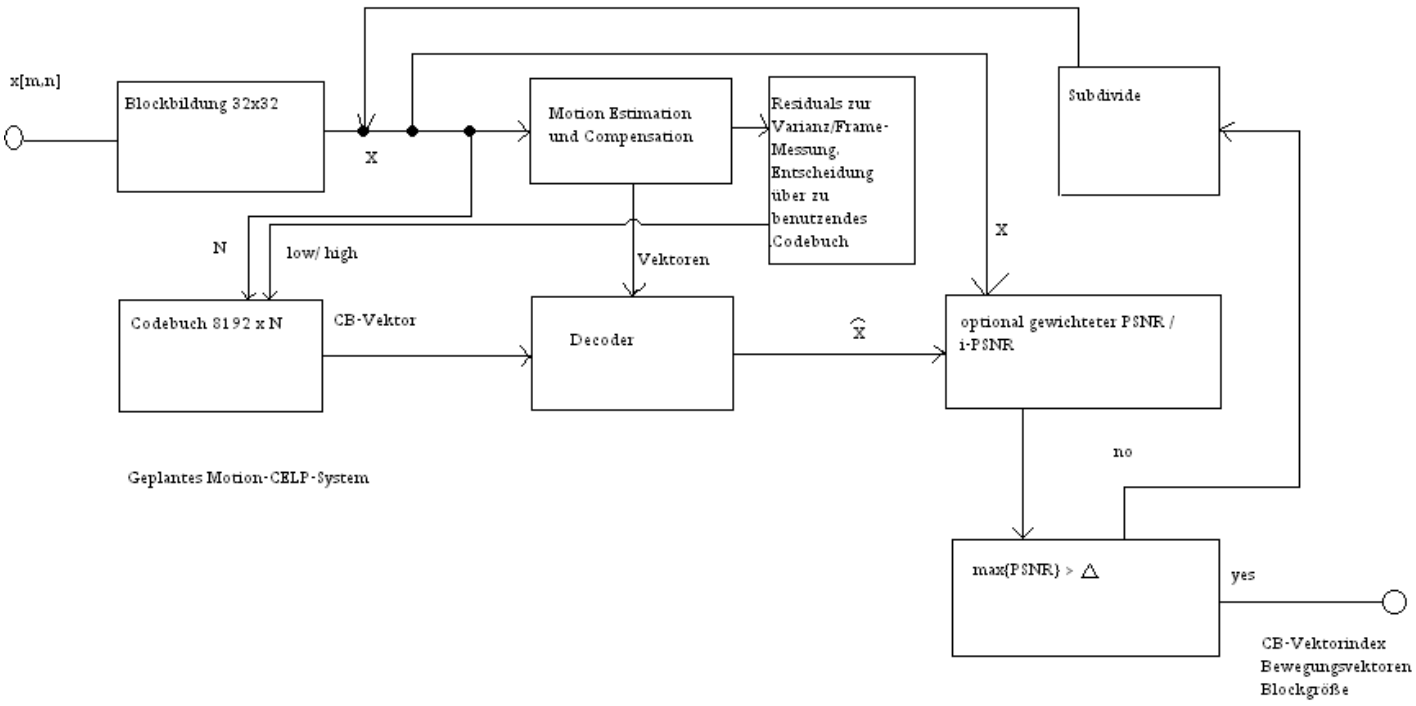
Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

5.] Das geplante Motion-CELP-System

Die Bewegungskompensation ist ein sehr mächtiges Werkzeug zur Erzeugung möglichst dekorrelierter Prädiktionsfehlersignale von Bildsequenzen und soll daher auch fester Bestandteil des Prewhitening-Filters im Motion-CELP-System sein.
Iterativ ausgewählte, zusätzliche spatiale Prädiktoren zur Entfernung von Rest-Korrelationen konnten in bisherigen Testläufen die Varianz zwar weiterhin um die Hälfte senken, zeigten jedoch hinsichtlich PSNR der dekodierten Frames, sowie des subjektivem Eindrucks keinerlei Verbesserung. Ihr Einsatz zusätzlich zur Bewegungskompensation ist daher als fraglich anzusehen, auch in Anbedacht, dass die für sie nötigen 3 Bit pro Block ebenfalls zu einer Verachtfachung des Codebuchumfangs genutzt werden könnten. Eine noch zu entscheidende Möglichkeit für ihren Einsatz (neben der I-Bild Verarbeitung) wäre, um Dekorrelation für Blöcke vorzunehmen, die durch Bewegungsschätzung nur schlecht vorhergesagt werden können, wobei ein zusätzlicher Flag gesendet werden müsste, ob der jeweilige Block temporal oder spatial prädiziert wurde.

Das System: Es wird jeweils ein Block mit 32 x 32 Pixeln eingelesen und eine möglichst genaue Prädiktion durchgeführt. Der jeweilig beste Prädiktor (gleich ob temporal oder spatial) bestimmt das Synthesefilter, welches prinzipiell mit jedem der im Codebuch abgelegten Vektoren angeregt wird. Eine PSNR-Messung mit dem Originalblock entscheidet über den jeweils besten Kandidaten. Liegt der maximal erzielte PSNR unterhalb eines bestimmten Thresholds, der als Qualitätsstufe einstellbar sein soll, so wird der Block in vier 16x16-Böcke geteilt und die Untersuchung beginnt von vorn. Ein Block wird solange aufgeteilt, bis der Threshold, oder aber die kleinstmögliche Blockgröße erreicht wird. Geplant sind 32, 16, 8 und 4er Blöcke.

Übertragen werden für jeden Block der jeweilige Codebuchvektorindex, die Blockgröße und der Bewegungsvektor (Prädiktorindex für I-Bilder). Betrachtet man alle 3 Kanäle, kommen für die Farbwerte nur die Codebuchindizes hinzu, da der Bewegungsvektor für alle Kanäle gleich ist und nur über den Helligkeitswert ermittelt wird.



Geplantes Motion-CELP-System

Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

6.] Bisherige Ergebnisse und Rückschlüsse

Bisherige Testläufe bezogen sich auf feste Blockgrößen von 32x32 - und 16x16 – Blöcken und dem 9.ten Frame aus „foreman.cif.yuv“, „silent.cif.yuv“ und „mobile.cif.yuv“. Das Codebuch für die 32er Blöcke wurden anhand Trainingsvektoren aus der foreman-Sequenz mittels LBG-Algorithmus auf einen Umfang von 8192 Einträgen gebracht, das 16er Codebuch auf 4096 Einträge.
Die Prädiktion wurde mittels Bewegungskompensation mit 8 Referenzbildern und ¼ - pel Genauigkeit durchgeführt.

Ergebnisse:

32-Block: foreman: 37,5 dB PSNR; silent: 44,5 dB PSNR; mobile: 27,9 dB PSNR
16-Block: foreman: 40 dB PSNR

Prozentual ergaben im foreman und silent – Frame viele der 32er Blöcke einen PSNR > 40 dB, diejenigen, die unterhalb dieser Schranke lagen , ergaben auch im 16er Block-Fall Schwierigkeiten. Um im Endsystem möglichst selten die Blöcke aufteilen zu müssen, sollten die Codebücher für jede Blockgröße gleichen Umfang aufweisen. Damit ergibt sich eine klare Genauigkeitserhöhung der Vektorquantisierung zusätzlich zur besseren Prädiktionsmöglichkeit, so dass der Threshold möglichst oft erreicht wird. Außerdem ist die Anzahl der Bit/Block damit für jede Blockgröße gleich und damit leicht handhabbar.

- Die zusätzliche Einbindung spatialer Prädiktoren erbrachte keine nennenswerte PSNR-Erhöhung im foreman-Fall.
- Auch wenn das Codebuch zu Testzwecken nur mit foreman-Trainingsvektoren gebildet wurde, ergab sich auch im Fall von „silent“ ein sehr hoher durchschnittlicher PSNR, bei „mobile“ jedoch einen kaum brauchbaren PSNR von 27,9 dB.

➔ Annahme: Betrachtet man die Residualvarianzen, so sind diese bei foreman und silent recht niedrig, bei mobile hingegen um ca. fünf mal höher.
➔ Rückschluss: Es wird je ein Codebuch für jede vorgesehene Blockgröße mit Trainingsvektoren aus verschiedenen Sequenzen einer gleichen Residual-Varianz-Klasse erstellt. Zunächst werden zwei solcher Klassen eingeplant, d.h. es wird ein Codebuch für niedrige Residualvarianzen und eines für höhere erstellt. Eine einfache Varianzmessung der Residuals während des Encodier-Vorganges entscheidet durch senden eines einzigen Bits am Ende eines Frames, oder nach mehreren Frames, welches der Codebücher verwendet wird.
➔ Eine andere Möglichkeit wäre eine adaptive Veränderung des Codebuches, wie in [Arya] beschrieben.

Weitere Tests müssen zeigen, ob im Falle der höheren Residualvarianzen der Codebuchumfang vergrößert werden sollte, oder aber eine zusätzliche spatiale Prädiktion hier doch gewinnbringend eingesetzt werden kann.

Bitratenbetrachtung:
Ohne nachträglicher Huffman- oder arithmetischer Codierung benötigt ein Block bei 8 Referenzbildern und 32er Integer-Suchweite im Bewegungskompensationsalgorithmus:

Für die Bewegungsvektoren: 8 + 8 Bit für x und y, 3 Bit für die z-Komponente (Referenzbild) = 19 Bit
Y-Codebuchindex: 13 Bit
U-Codebuchindex: 12 Bit [halber Codebuchumfang für Farbwerte]
V-Codebuchindex: 12 Bit
Ges.: 56 Bit/Block für alle 3 Kanäle

Für den reinen 32er – Block – Fall ergibt sich damit ein Compressionsfaktor von Cf_{32} = [32*32*3*8 / 56] = 438,85.
Für den reinen 16er-Block-Fall entsprechend ein Viertel davon, also Cf_{16} = 109,7.

Die zusätzlichen 2 Bit/Block im endgültigen System zur Blockgrößenangabe wurden nicht betrachtet.

In Anbetracht dessen, dass im H.264-System bereits ein sehr wirksames Konzept besteht, die Bewegungsvektoren prädiktiv zu codieren, indem benachbarte Vektoren genutzt werden und die resultierenden Differenzwerte Huffmancodiert werden, und der Annahme, dass sich benachbarte Blöcke ähneln, ist die Wahrscheinlichkeit hoch, dass sich auch, bei geeigneter Ordnung des Codebuches, die Indizes auf gleiche Weise prädiktiv codieren lassen.
Weiterhin kann angenommen werden, dass die U und V-Blöcke zumeist von höherer Qualität sind. Für sie wird ein gemeinschaftliches Flag gesetzt, d.h. ihre Unterteilung in Subblöcke ist von der der Helligkeitsblöcke entkoppelt.

Im Resultat wird eine Bitmenge von etwa 35-38 Bit/Block erwartet, bzw. angestrebt. [Mit Bauchschlag]

Somit ergäbe sich: Cf_{32} = 646 bis 702.

Bei „foreman.cif.yuv“ mit 10 Hz Bildrate ergäbe sich: R_{32} = 37,62 kbit/s bis 34,65 kbit/s. Kann dabei der anhand des 9.Frames gemessene durchschnittliche PSNR von 37,5 dB eingehalten werden, so wäre dieses Ergebnis in Gegenüberstellung von veröffentlichten H.264 – Messungen der gleichen Test-Sequenz superior.
Diese Vergleichs-Ergebnisse sind jedoch uneinheitlich, da im H.264-Standard der Encoder nicht festgeschrieben ist. Eine derartige Qualität ergab sich jedoch aus letzten Recherchen erst ab einer Bitrate von R_{H264} > 150 kbit/s.

In [Arya] wurde 2008 ein System zur Übertragung von Teleconferencing vorgeschlagen, das mittels dreidimensionaler Vektorquantisierung bei der „Miss America“-Sequenz ein PSNR von 32 dB und Cf von 225 erlangt. Ausgehend der letzten Testläufe ist anzunehmen, dass das Motion-CELP-System eine bessere Performance bieten wird.

Eine weitergehende Forschung wäre also durchaus sinnvoll.

7.] Weitere geplante Implementierungen:

- Zur Bekämpfung von Blockartefakten, welche durch die Bewegungskompensation herrühren, wird ein Deblocking-Filter, wie es H.264 nutzt getestet.
- Ein wichtiger Aspekt in der CELP-Sprachcodierung ist die psychoakustische Frequenzgewichtung des Fehlersignals. Wurde dies in [Dubois2] vernachlässigt, soll zum Abschluss der Arbeit ein Schalter implementiert werden, der dem Benutzer erlaubt zwischen optimaler objektiver und subjektiver Qualität zu entscheiden.
  Im letzten Fall soll neben der Untersuchung von wahrnehmungsbasierter Frequenzgewichtung des Fehlers, auch untersucht werden, inwieweit zur Zeit bekannte Modelle zur Bestimmung visueller Aufmerksamkeit [ROI] in das System eingebunden werden können, um z.B. für Nicht-ROI-Areale einen niederen PSNR-Threshold zuzulassen. Weiterhin soll ein inhaltsbasierter PSNR [i-PSNR] vorgeschlagen werden, derart, dass Fehler in Kantennähe stärker gewichtet werden, da das HVS [Human Visual System] Strukturen stärker wahrnimmt, als Texturen. Auf umfassende empirische Analysen zur exakten Definition eines solchen i-PSNR wird aus Umfang-Gründen innerhalb der Arbeit verzichtet. Eine denkbare Definition wäre, dass ein i-PSNR von k dB eines Kanten-beinhaltenden Blocks, die gleiche subjektive Qualität zeigen soll, wie ein kantenloser Block mit (herkömmlichen) PSNR von k dB.

Andreas Tölgyesy, Matr.Nr.:198355, TU-Berlin
Studiengang: Kommunikationswissenschaft / Technische Akustik
Email: tijuanalife@web.de ; Tel: 01716282340

8.] Nächste Schritte

Geplante nächste Testläufe: Parallel zur Implementierung des End-Systems, werden als nächste Schritte weitgehende Testläufe mit verschiedenen Sequenzen, über jeweils viele Frames mit fester 32er. Blockgröße gestartet. Dabei werden zunächst nur die Y-Werte betrachtet, sowie die Ergebnisse ab dem 9. Frame, damit 8 Referenzbilder zur Bewegungskompensation genutzt werden können. Ziel dabei ist, sich einen vertrauenswürdigeren Überblick über zu erwartende Ergebnisse zu verschaffen.
Testläufe mit festen 16er. Blöcken, sowie alsbald mit dem endgültigen blockadaptiven, residual-varianz-adaptiven System werden folgen.
Zum Abschluss sollen dann weitgehende Testläufe mit dem zusätzlich HVS-optimierten System durchgeführt werden.

Referenzen:

[Shu]: Shu,Y. ; Robinson J.A. : Robust Motion Picture Residue Coding for Noisy Channels; 7th Int. Conf. on Image Processing and its applications, (IPA'99,IEE No. 465), Manchester,UK,13-15 July 1999, 153-156

[Strutz]: Thilo Strutz: Bilddatenkompression, 3. Auflage, Vieweg

[Dubois]: H. Nguyen ; E. Dubois : A two-dimensional code excited linear prediction (2 D-CELP) image coding system; in Image Processing Algorithms Techniques, Proc. SPIE 1244, Feb. 1990, pp 190-198

[Dubois2]: Sonia Aissa ; E. Dubois : 2-D-CELP Image Coding with Block-Adaptive Prediction and Variable Code-Vector Size ; IEEE Transactions on Image Processing, Vol. 5, No. 2, Feb. 1996

[Arya]: V. Arya; A. Mittal; A. Pande; R.C. Joshi: An Efficient Coding Methoid for Teleconferencing Video and Confocal Microscopic Image Sequences; Journal of Computing and Information Technology – CIT 16, 2008,3, 145-156