

# Qualitätsmessung

## Übersicht

Einleitung  
Grundlagen  
MSE  
PSNR  
SSIM  
Quellen

## Einleitung

Bei der verlustbehafteten Kompression von Videodaten unterscheiden sich Ausgangsbild und Ergebnis. Es wird versucht, Bildinformationen, die das visuelle System des Menschen nicht oder nicht so stark wahrnimmt, zu entfernen oder mit geringerer Genauigkeit abzuspeichern. Je größer die Kompression, desto mehr Informationen müssen entfernt werden. Bei zu großer Kompression wird der Unterschied zur Quelle immer stärker sichtbar, die „Qualität“ leidet. Zur Messung der Qualität von Videomaterial gibt es verschiedene Verfahren, die einen Zahlenwert ergeben, der etwas über die Qualität des Materials aussagt.

Man kann zum einen die mathematischen Unterschiede zwischen dem ursprünglichem und dem komprimierten Videomaterial messen. So kann man erkennen, wie viele Bildinformationen entfernt wurden. Da jedoch die menschliche Wahrnehmung bestimmte Informationen nicht oder nicht so stark wahrnimmt, sagt die rein mathematische Ähnlichkeit nicht viel über die tatsächliche wahrgenommene Bildqualität aus.

Andere Verfahren versuchen, das menschliche visuelle Wahrnehmungssystem mit mathematischen Modellen nachzuempfinden und so die wahrgenommene Qualität zu errechnen.

Wir werden hier die 3 populärsten Qualitätsindizes unter die Lupe nehmen:

- MSE (Mean Squared Error)
- PSNR (Peak Signal to Noise Ratio)
- SSIM (Structural SIMilarity)

## MSE

Der MSE (Mean Squared Error) beschränkt sich auf die rein qualitative Betrachtung von Bildfehlern. Ein Referenz und ein Testbild werden miteinander verglichen, und es wird für jedes Pixel der Unterschied in der Helligkeitskomponente betrachtet und diese Differenzen quadriert und aufsummiert. Diese Summe wird dann über die Anzahl der Pixel gemittelt. Als Ergebnis erhalten wir die durchschnittliche quadrierte Abweichung zwischen den beiden Bildern.

Manchmal wird anstelle des MSE auch der RMSE (Root Mean Squared Error) angegeben. RMSE ist die Quadratwurzel aus MSE.

## PSNR

PSNR (Peak Signal to Noise Ratio) basiert auf dem MSE und ist im Grunde eine Anpassung an die menschliche Wahrnehmung, die Reize logarithmisch wahrnimmt.

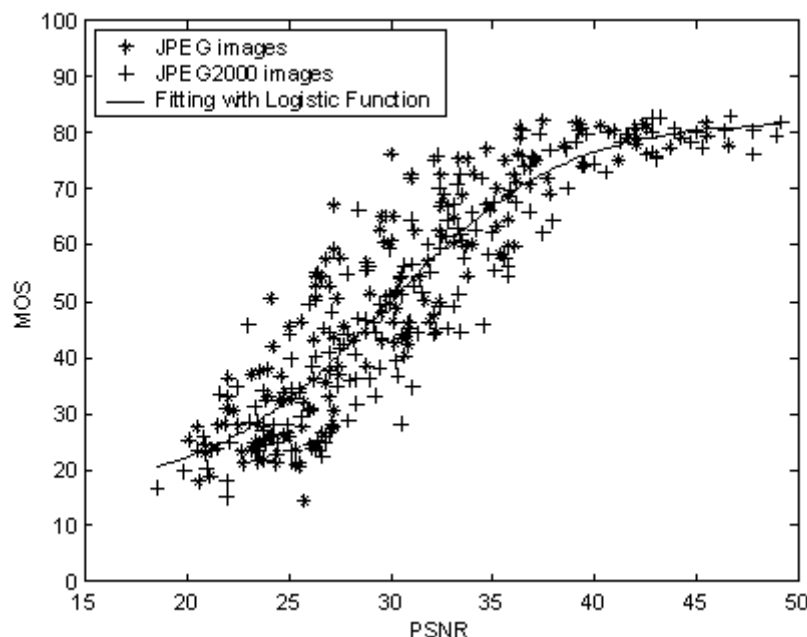
PSNR (Einheit dB) ist folgendermaßen definiert:

$$PSNR = 20 \cdot \log\left(\frac{255}{RMSE}\right)$$

Es gibt einige leicht davon abweichende Definitionen, diese ist aber die gebräuchlichste.

Je größer der RMSE, desto kleiner der PSNR Wert. Also gilt bei PSNR: Je größer desto besser.

Wenn man viele Test und Referenzbilder von einer Anzahl Testpersonen bewerten lässt und die gemittelte Wertung gegen den PSNR Wert aufträgt, erhält man folgendes Diagramm [1]:









Auf der y-Achse ist die Bewertung der Testpersonen aufgetragen (MOS, Mean Opinion Score), auf der x-Achse stehen die PSNR Werte.

Wenn man alle Markierungen bei einem bestimmten PSNR Wert betrachtet, so sieht man vor allem im Mittelteil des Graphen eine enorme Spannweite. Bilder mit einem PSNR von ungefähr 30 dB wurden von den Testpersonen mit Wertungen zwischen 25 und fast 80 auf einer Skala zwischen 1 und 100 gesehen. Bilder, die von den Testpersonen mit einer Qualität von 80 bewertet wurden, hatten PSNR Werte zwischen 30 und 50 dB.

Diese enorme Diskrepanz zwischen der wahrgenommenen und der mathematischen Qualität ist die große Schwäche von MSE und PSNR. Das kann man auch sehr gut an dieser Bilderserie verdeutlichen:

Hier einige Beispielbilder [1], alle haben einen identischen MSE und somit auch PSNR:

		
Referenzbild SSIM = 1.0000	Kontrastkorrigiertes Bild SSIM = 0.9168	Helligkeitskorrigiertes Bild SSIM = 0.9900
		
Extreme JPEG Blockartefakte SSIM = 0.6949	Weichgezeichnetes Bild SSIM = 0.7052	„Salz und Pfeffer“-Rauschen SSIM = 0.7748

Obwohl diese Bilder konstuiert sind und in der Realität nicht in dieser extremen Form auftreten, zeigen sie doch die Schwächen dieser beiden Verfahren klar auf.

PSNR ist in der compare() Funktion von Avisynth eingebaut, dort werden auch noch einige weitere statistische Daten über die Unterschiede der 2 Clips, die der Funktion übergeben werden angezeigt.

## SSIM

Einen ganz anderen Ansatz verfolgt der Qualitätsindex SSIM. Hier wird versucht, die Ähnlichkeit der Bildstruktur (Structural SIMilarity) zu erfassen.

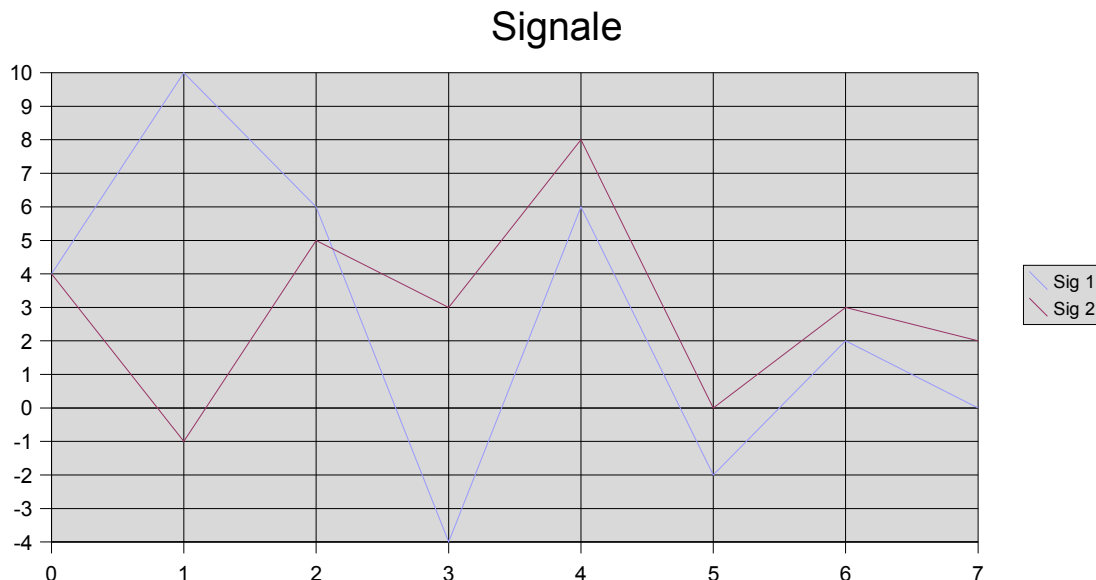
Der Begriff „Struktur“ eines Bildes bezeichnet die Abhängigkeiten zwischen benachbarten Bildpunkten, die unabhängig von der Helligkeit und dem Kontrast in dieser Region des Bildes sind.

Beim Vergleich zweier Bilder (Videoframes) wird ein Bildfenster von 11x11 Pixeln (dieses Fenster wird einer drehsymmetrischen Gaussverteilung gewichtet, um weiter entfernte Pixel weniger stark zu berücksichtigen) nacheinander über das ganze Frame bewegt, so dass jeder Bildpunkt einmal in der Mitte des Fensters war. Jetzt werden die Inhalte beider Fenster verglichen. Das geschieht in 3 Stufen:

- Der Erste Test (luminance comparison) untersucht, ob die Bildfenster sich in der durchschnittlichen Helligkeit unterscheiden.
- Der zweite Test (contrast comparison) untersucht, wie groß der Unterschied zwischen den Kontrasten der Fenster ist.
- Im dritten Test (structural comparison) wird die Struktur der Bilder verglichen. Jeder dieser Test ergibt eine Maßzahl, die zwischen 0 (total unterschiedlich) und 1 (identisch) liegt.

Die Maßzahlen werden verrechnet. Dann wird das Bildfenster um ein Pixel verschoben und das Verfahren für das neue Fenster wiederholt. Am Ende werden die Wertungen für alle Fenster zu einer Gesamtwertung zusammengefasst, oder es wird anhand der Werte eine Fehlerkarte für das Bild erstellt.

Um den Ansatz zu verdeutlichen, betrachten wir jetzt 2 Signale, die wir vergleichen wollen. Diese zwei Signale seien Teile der Helligkeitskomponente einer Bildzeile, das blaue Signal ist das Original und das rote Signal ist z.B. durch verlustbehaftete Kompression verändert.

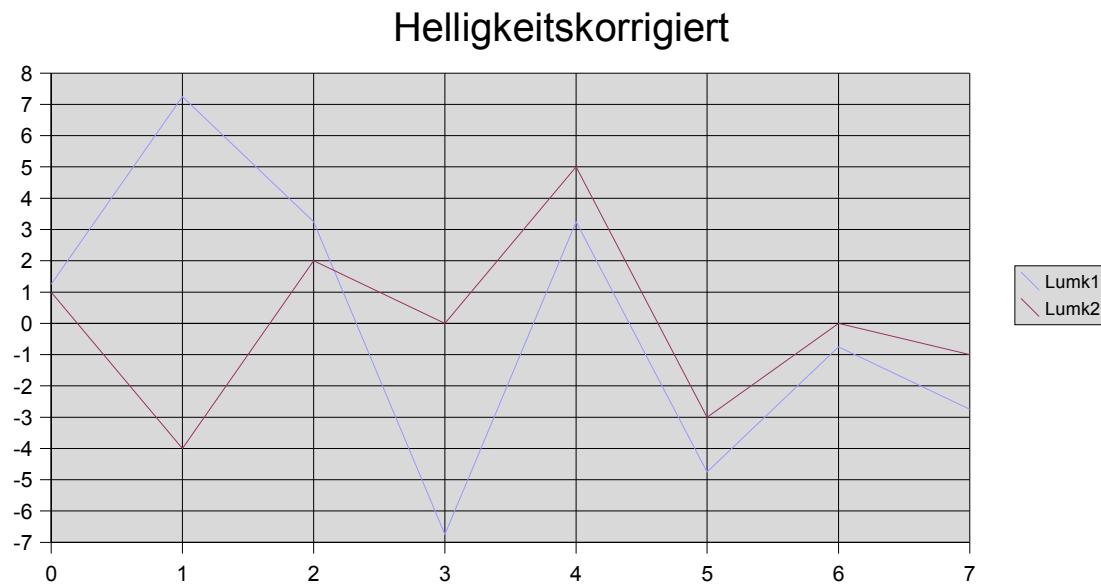


Die negativen Werte im blauen Signal machen zwar nicht wirklich Sinn, aber es geht hier in erster Linie ums Prinzip.

Im ersten Schritt wird die lokale Bildhelligkeit bestimmt, also die Helligkeit des Bildes an der betrachteten Stelle. Da hohe Werte für helle Pixel stehen und niedrige für dunkle kann man die Durchschnittshelligkeit als Durchschnitt der Pixelwerte berechnen. Die Durchschnittshelligkeiten sind als grüne Linie für das rote Signal und als lilane Linie für das blaue Signal in das Schaubild eingetragen.

Die beiden Bildteile haben unterschiedliche Bildhelligkeiten. Diese unterschiedlichen Helligkeiten fließen in eine erste Wertung, dem Helligkeitsvergleich (luminance comparison) ein.

Als nächstes werden die beiden Bildteile auf die gleiche Durchschnittshelligkeit gebracht, indem die jeweilige durchschnittliche Helligkeit abgezogen wird:

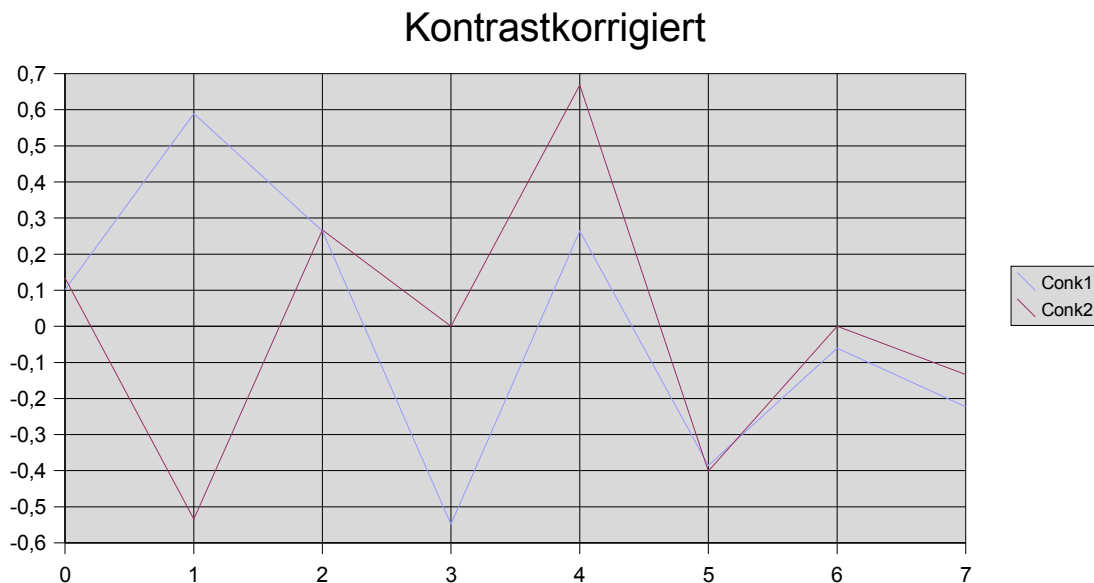


Jetzt wird von den helligkeitskorrigierten Signalen der Kontrast bestimmt. Als Maß dafür nimmt man die sog. Standardabweichung, die Wurzel des mittleren Quadrats der Abweichung vom Mittelwert der Helligkeit.

Diese Standardabweichung ist umso größer, je größer die Ausschläge des Signals sind, d.h. Je unterschiedlicher hell die einzelnen Pixel unserer Bildzeile sind. Das rote Signal hat also eine größere Standardabweichung als das blaue Signal, und somit einen größeren Kontrast.

Die unterschiedlichen Kontraste fließen in die zweite Wertung, den Kontrastvergleich oder contrast comparison ein.

Die Signale werden jetzt auch auf die selbe Standardabweichung und somit Kontrast gebracht, indem sie durch ihre Standardabweichung geteilt werden.

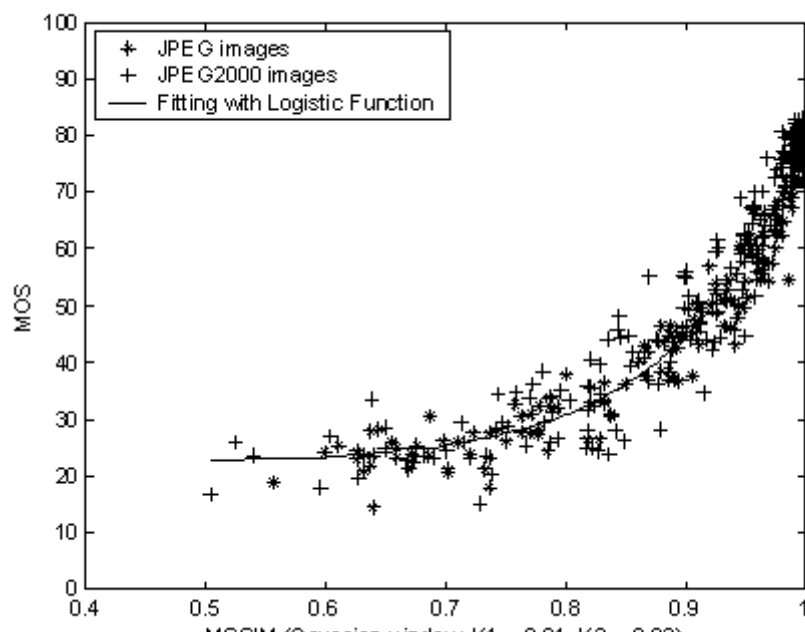


Um die Ähnlichkeit zwischen diesen beiden Signalen zu bestimmen, bedient man sich eines mathematischen Werkzeugs, dem Skalarprodukt oder auch innerem Produkt. Das Ergebnis des Skalarprodukts der beiden Vektoren fließt in die dritte Wertung ein, den Strukturvergleich (structural comparison).

Die drei Wertungen werden kombiniert und ergeben den SSIM Wert für diese beiden Bildteile, der zwischen 0 und 1 liegt. Je höher die Werte, desto größer die strukturelle Ähnlichkeit.

SSIM ist gegenüber Helligkeits- und Kontraständerungen nicht so empfindlich wie PSNR, da diese Faktoren separat bewertet werden. Das deckt sich mit unserer Wahrnehmung, die Helligkeits- und Kontrastverschobenen Bilder oben empfinden wir qualitativ nicht als wesentlich schlechter, während die Bilder, bei denen die Bildstruktur zerstört ist, deutlich schlechter aussehen.

Auch statistische Messungen haben ergeben, dass SSIM den subjektiven Qualitätseindruck deutlich besser modelliert als PSNR.



Man sieht auf diesem Diagramm [1], dass die Bewertungen der Testpersonen und die SSIM Bewertungen viel weniger weit auseinanderliegen, als die PSNR. Der MOS für Bilder mit gleichem SSIM Wert unterscheiden sich in Extremfällen um 30 Punkte, im Normalfall aber nur zwischen 10 und 20 Punkte.

Die SSIM Werte für Bilder gleichen MOS liegen höchstens ca 0.35 in den niederqualitativen Bereichen auseinander.

Generell gilt: Je besser die Qualität der Bilder, desto mehr stimmen SSIM und MOS übereinander.

Jedoch sieht man, dass trotz der besseren Korrelation, nicht annähernd von einer objektiven Qualitätsmessung gesprochen werden kann. Man sollte deshalb SSIM nicht zu hoch bewerten.

Es gibt ein SSIM Plugin von lefungus für avisynth, das es erlaubt, mit avisynth SSIM Messungen durchzuführen.

Jedoch weicht die Implementierung aus Geschwindigkeitsgründen vom ursprünglichen Konzept ab. Es wird ein 8x8 Bildfenster ohne Gewichtung verwendet, ausserdem wurde ein Lumimasking Mechanismus eingebaut, der die SSIM Wertungen nochmals nach psychovisuellen Gesichtspunkten gewichtet. Der Originalentwurf sieht weiterhin nur die Bewertung der Helligkeitswerte vor, während das Plugin auch die Farbigkeit strukturell auswertet.

## Quellen:

[1]<http://www.cns.nyu.edu/~zwang/files/research/ssim/index.html>

(Inzwischen ist die endgültige Version des Papers über SSIM verfügbar, die Informationen über SSIM hier stammen aus einem Vordruck vom Mai 2004)