COMPRESSIONE DIGITALE

Appunti
sulle tecniche
di
compressione Audio e Video

a cura di
Antonio Silvestri

Febbraio 2003

INDICE

CAPITOLO 1

INTRODUZIONE

Col termine compressione digitale oppure compressione dati si intendono tutte quelle tecniche usate per ridurre la quantità di dati per rappresentare una determinata informazione. Un esempio sono i *file* di tipo **jpeg** o **gif** i quali contengono valori che rappresentano delle immagini.
Tali tecniche, in generale, sfruttano il fatto che l'informazione non ha un carattere casuale ma contiene un certo grado di ordine. Se si è nelle condizioni di poter estrarre e duplicare questo ordine, allora l'informazione potrà essere rappresentata con meno dati rispetto all'originale e sarà possibile ricostruire, in un secondo tempo, l'informazione iniziale, oppure, a nostra scelta ed in dipendenza della tecnica adottata, una sua fedele approssimazione.
Sono moltissime gli impieghi di queste tecniche, e non solo nel campo multimediale. Si pensi, ad esempio, agli algoritmi di compressione dei programmi di utilità *winzip*, *compress*, *gzip*, *winrar*, ecc. molto usati sia direttamente che indirettamente da molte applicazioni.
Naturalmente l'area applicativa primaria è quella del Video e dell'Audio digitale, ove le motivazioni dell'uso di queste tecniche risultano evidenti se si prendono in considerazione le grandezze in gioco quando si devono manipolare, trasmettere ed elaborare dati di tipo audiovisivo. Ecco alcuni esempi:

- L'audio stereo digitale di un CD è campionato a 44.100 Hz e quantizzato con 16 bits/campione, questo corrisponde ad un *birate* (numero di bit al secondo) di 1,345 Mbits/s (44100 **x** 16 **x** 2 canali = 1411200 bits). Quindi in un minuto vengono generati circa 10 MB di dati.

- Una sequenza video della durata di un minuto può generare quasi 1,2 GB con un bitrate di circa 158[1] Mbits/s.

- Una sequenza video HDTV (TV ad alta definizione 1920 **x** 1152 pixels) *widescreen* (16:9) genera un bitrate di circa 1 Gbits/s, quindi in un minuto avremo una quantità di dati superiore ai 6 GB.

È palese la necessità di comprimere, ridurre il bitrate di questi flussi, in certi casi enormi, di dati; questa area di ricerca è molto attiva per l'impatto che ha e che avrà ancora di più in futuro.

---

[1] $[\underbrace{(720 \times 576)}_{\text{pixel luminanza}} + \underbrace{2 \times (360 \times 576)}_{\text{colore (formato 4:2:2)}}] \times \underbrace{25}_{\text{frame al secondo}} \times \underbrace{8}_{\text{bits per byte}} = 165888000 \; bits / s$

VIDEO ANALOGICO (TV)

La trasmissione di una ripresa televisiva avviene con la conversione di ogni elemento dell'immagine, pixel[2], in una rappresentazione elettrica per mezzo di un apparato a scansione con celle fotosensibili (telecamera analogica o digitale). La scansione è effettuata riga per riga da sinistra a destra e dal basso verso l'alto, il segnale così ottenuto modula una frequenza portante per la successiva fase di trasmissione. Affinché l'immagine sia fedelmente ricostruita dal tubo catodico (o da altro sistema analogo) del televisore, il pennello elettronico deve essere sincronizzato con quello della scansione; ciò è assicurato dai segnali di sincronismo orizzontale e verticale aggiunti in fase di trasmissione.

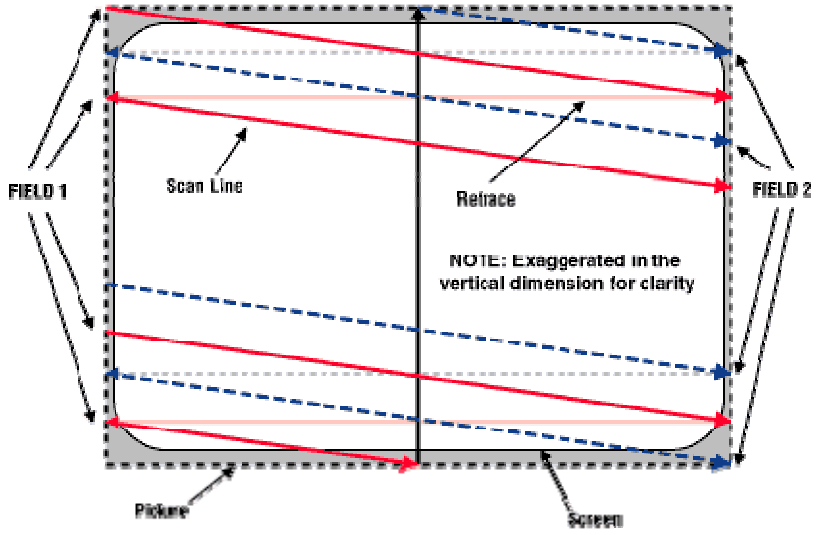


Figura 1: immagine video interlacciata

Negli attuali standard la scansione e riproduzione delle righe che compongono un'immagine televisiva (detta quadro o *frame*) non sono effettuate riga per riga (metodo chiamato progressivo) ma sono formate da due campi (detti semiquadri o *fields*) interlacciati (Figura 1). Ogni campo consiste dalle righe pari dell'immagine e l'altro da quelle dispari. Questo ingegnoso sistema fu realizzato da una parte per diminuire lo sfarfallio[3] (*flicker*) dovuto alla bassa frequenza di quadro (*refresh rate*) e dall'altro per non aumentare la banda passante del segnale video.

TV a colori

L'introduzione del colore ha reso necessario lo studio di un sistema compatibile con i ricevitori in bianco e nero. Bisognava veicolare il segnale relativo al colore, detto di *crominanza*, nella stessa banda del segnale monocromatico, detto di *luminanza*.
Come colori primari per sintetizzare tutti gli altri furono scelti il Rosso, il Verde ed Blu (RGB), mentre come segnale di luminanza fu individuata una loro opportuna combinazione lineare che tenesse conto delle caratteristiche percettive dell'occhio umano. Sono stati quindi definiti tre tipi di segnali, allo scopo di trasmettere il colore in modo compatibile col sistema moncromatico, essi sono:

---

[2] Picture Element

[3] Tale condizione avviene poichè l'illuminazione ambientale durante la visione dei programmi televisivi è abbastanza elevata da rendere insuffuciente un frame rate di 25 Hz.

$$Y = 0.299R + 0.587G + 0.114B$$
$$R\text{-}Y = 0.701R - 0.587G - 0.114B$$
$$B\text{-}Y = -0.299R - 0.587G + 0.886B$$

Y trasporta il segnale relativo alla parte compatibile col segnale bianco e nero. L'informazione del colore è affidata alle due componenti R-Y e B-Y chiamati segnali *differenza colore*, il segnale differenza colore del verde $G-Y$ resta determinato dagli altri e da Y .
Per trasmettere i suddetti segnali nella stessa banda di quello di luminanza Y viene usato un sistema ad "incastro" (*frequency interleaving*). Questo procedimento sfrutta la distribuzione spettrale regolare del segnale video monocromatico, che presenta un addensamento di righe intorno alle frequenze multiple di riga (configurazione chiamata a pettine).

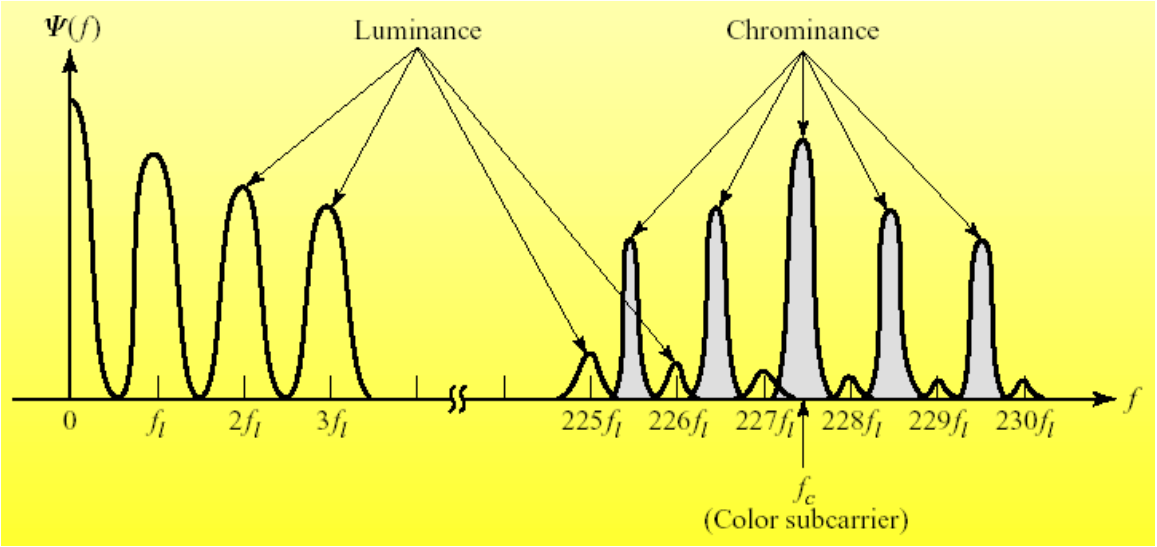


Figura 2: Frequency Interleaving. f è la frequenza di riga

I segnali R-Y e B-Y che a meno di un fattore di scala vengono anche indicati con U, V o C_b, C_r modulano, con una tecnica chiamata "modulazione di ampiezza in quadratura" o **QAM** dall'inglese "**Quadrature Amplitude Modulation**", una stessa frequenza ausiliaria chiamata sottoportante di crominanza scelta in modo che risulti un multiplo dispari della frequenza di riga[4]. Poiché la distribuzione spettrale di tale segnale è analoga a quello di luminanza, si avrà come risultato che le righe dello spettro del segnale di crominanza si posizioneranno tra le righe dello spettro di quello luminanza (Figura 2).

[4] Con questa scelta il segnale di crominanza si posiziona nella parte alta dello spettro di quello di luminanza ove tale segnale ha un ampiezza piccola. In questo modo qualsiasi interferenza tra i due segnali è ridotta al minimo.

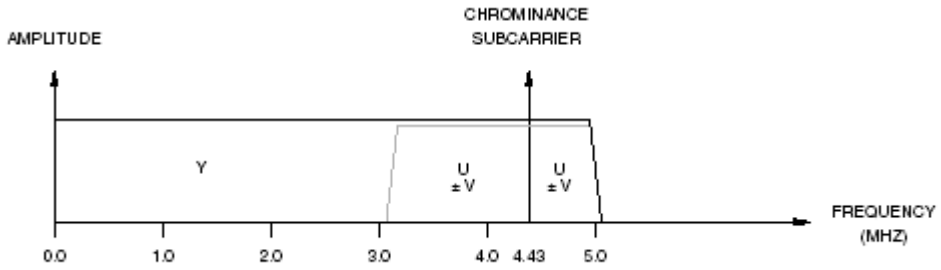


Figura 3: Spettro del segnale Video

Con questo metodo risulta che il colore è associato alla fase del segnale di crominanza mentre la sua saturazione all'ampiezza del medesimo. Nella Figura 3 è mostrato lo spettro completo del segnale video PAL.

TV Standards

Attualmente esistono tre standards televisivi a colori l'NTSC (National Television Standard Committee) usato principalmente negli USA, il PAL (Phase Alternating Line) ed il SECAM (Séquentiel Couleur à Mémoire) usati in Europa (Figura 4).
Il SECAM ed il PAL sono derivati dall'NTSC e ne correggono il principale difetto: L'instabilità cromatica dovuta ad errori di fase del segnale di crominanza durante la trasmissione. Per ovviare a questo inconveniente che produce colori errati nel ricevitore, entrambi i sistemi partono dal presupposto che l'eventuale errore di fase subito dal segnale influenzi in maniera praticamente uguale due righe consecutive.
Per il sistema PAL, la tecnica prevede di invertire, in trasmissione, la fase del segnale di crominanza a righe alterne. Nel ricevitore la fase viene ristabilita e mediando con la riga precedente l'eventuale errore viene cancellato.
Il sistema SECAM risolve il problema trasmettendo i due segnali differenza colore in modulazione di frequenza (FM) ed una alla volta su due righe consecutive. È cura del ricevitore ricombinare insieme i due segnali.

Parameters	NTSC	PAL	SECAM
Field Rate (Hz)	59.95 (60)	50	50
Line Number/Frame	525	625	625
Line Rate (Line/s)	15,750	15,625	15,625
Color Coordinate	YIQ	YUV	YDbDr
Luminance Bandwidth (MHz)	4.2	5.0/5.5	6.0
Chrominance Bandwidth (MHz)	1.5(I)/0.5(Q)	1.3(U,V)	1.0 (U,V)
Color Subcarrier (MHz)	3.58	4.43	4.25(Db),4.41(Dr)
Color Modulation	QAM	QAM	FM
Audio Subcarrier	4.5	5.5/6.0	6.5
Total Bandwidth (MHz)	6.0	7.0/8.0	8.0

Figura 4: Specifiche dei principali standard TV

VIDEO DIGITALE

Se si vuole digitalizzare un segnale analogico con una banda passante F_{max} è necessario campionare tale segnale ad una frequenza F_{camp} almeno doppia della massima frequenza presente nel segnale analogico (teorema di Shannon/Nyquist). Questo per evitare errori (***aliasing***) nella successiva fase di ricostruzione del segnale analogico.
Per il campionamento del segnale video digitale lo standard ITU-R BT.601[5] prescrive che il segnale di luminanza Y sia campionato con una frequenza di F_{camp} = 13,5 MHz e quello di crominanza a F_{camp} = 6,75 MHz. Da ciò risulta che ci saranno 720 campioni per riga relativi alla luminanza e 360 campioni per riga per ognuno dei due segnali differenza colore ed un totale di 576 righe per il sistema PAL e 480 per quello NTSC. A tale formato si fa riferimento come *campionamento 4:2:2*.

La sigla `x:y:z`, che si incontra spesso nel settore video digitale ha il seguente significato: Per ogni **x** campioni di luminanza vi sono **y** campioni di crominanza per ogni riga alterna (righe dispari) e **z** per tutte le altre (righe pari). In pratica il formato in componenti YUV 4:2:2 (Figura 5) indica che per ogni 4 pixel relativi alla luminanza ve ne sono due relativi al colore, per tutte le righe; tale rappresentazione ha come conseguenza una compressione di 1,5:1 alla sorgente. Questo è il formato usato nel mondo del video professionale ed in quello delle trasmissioni analogiche/digitali (TV terrestre e satellitare).
Il formato 4:2:0 (Figura 6) che descrive un campionamento del colore della metà sia in senso orizzontale che verticale viene usato negli standard di compressione MPEG (DVD). Tale formato produce in partenza una compressione 2:1.

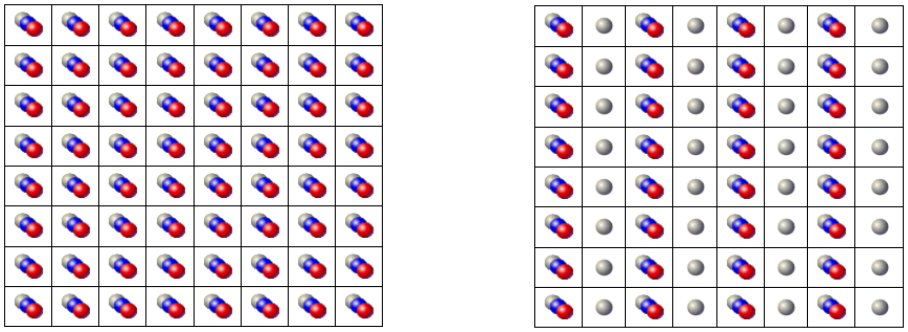


Figura 5: Campionamento 4:4:4 e 4:2:2

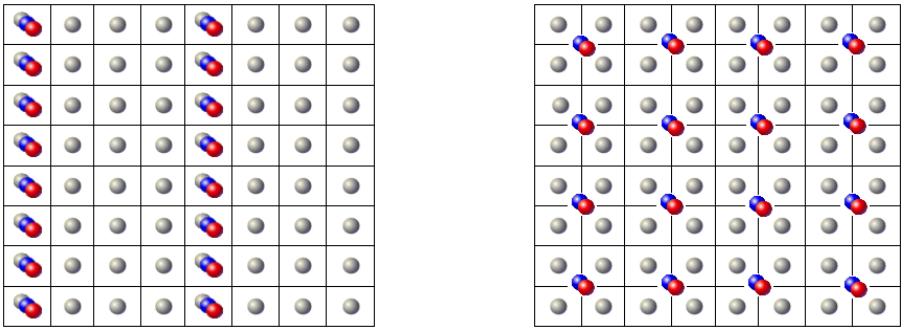


Figura 6: Campionamento 4:1:1 (a destra) e 4:2:0 (8x8 pixel per la luminanza e 4x4 pixel per la crominanza)

[5] Questo è la nuova sigla del CCIR il quale era il settore delle radiocomunicazioni del CCITT (Comité Consultatif International du Télégraphe et du Téléphone), di recente rinominato ITU (International Telecomnunications Union)

La rappresentazione matematica dello **spazio di colore** in componenti YUV risulta vantaggioso rispetto a quello RGB per varie ragioni, una di queste è la velocità: Nel formato 4:2:0 ci sono metà dei pixel per frame, il quale implica una maggiore rapidità in tutte le manipolazioni al quale può essere sottoposto. Un'altra ragione è la limitata sensibilità dell'occhio umano per la perdita di informazione sul colore.
Inoltre molte operazioni di elaborazione su di un'immagine, come ad esempio la modifica della luminosità o del colore, sono molto più efficienti se fatte separatamente sui componenti YUV invece di essere applicate contemporaneamente su tutte e tre canali RGB. Un altro motivo è una migliore qualità dell'immagine. Il ***gamut*** [6] dello spazio di colore RGB è inferiore a quello a componenti YUV. Quindi, di solito, non è una buona idea fare conversioni da YUV → RGB poiché implica una perdita di informazione.

[6] Il numero dei colori che una data rappresentazione (spazio di colore) può generare.

CAPITOLO 2

TECNICHE DI COMPRESSIONE: CLASSIFICAZIONE

Le rappresentazioni digitali delle immagini, delle sequenze video e dell'audio possiedono una considerevole ridondanza statistica di valori. All'interno delle immagini o del singolo video-frame (fotogramma) esiste una correlazione di tipo spaziale al quale si fa riferimento con il termine di *ridondanze spaziali*. Allo stesso modo si definiscono *ridondanze temporali* la correlazione presente tra i frame (fotogrammi adiacenti) che compongono un video.
Inoltre dal punto di vista percettivo molte informazioni sono irrilevanti. È quindi possibile sfruttare tutte queste ridondanze e irrilevanze per ridurre la quantità di dati.
Le tecnologie di compressione sono divise in due grandi famiglie: quelle senza perdita di informazione chiamate anche **lossless** e quelle di tipo **lossy** nelle quali è permessa una perdita, controllata, di informazione.
Tutte queste tecnologie devono assicurare che i dati compressi, rappresentativi di una determinata informazione, veicolino solo quelle informazioni percettibili dal destinatario.

Lossless Compression

La compressione è reversibile o *lossless* quando il procedimento usato permette il recupero completo dell'informazione originaria così come era prima che venisse applicata l'algoritmo di compressione. In tali sistemi il rapporto di compressione che si può ottenere è caratterizzato da una grande variabilità dipendente dal tipo di dati ed è in genere modesto (2:1, 3:1). Esempi di tecniche di compressione *lossless* di nostro interesse sono:

- **Entropy Encoding** o codifica entropica, anche chiamata codifica Huffman (*Huffman Coding*) o codifica a lunghezza variabile o con la sigla **VLC** dall'inglese *Variable Length Coding*. Questa è un tipo di codifica nella quale vengono associati ai valori da codificare probabilità distinte in base alla loro distribuzione statistica. Ai valori più probabili viene assegnato un codice breve mentre, ai rimanenti, codici sempre più lunghi.

- **Run Length Encoding** (**RLE**). In genere i dati che rappresentano una qualunque informazione (un'immagine, un programma, una sequenza video o audio) mostrano delle sequenze di valori ripetuti. A quei valori ripetuti si sostituisce la coppia <*numero di ripetizioni*>, <*valore*>.

- **Codifica Predittiva** (*Predictive Coding*): Codifica a previsione nella quale vengono trasmessi solo differenze rispetto ad un valore di riferimento. Questa tecnica, di per sé, non produce compressione, ma opera una trasformazione sui dati, tale da rendere vantaggioso l'applicazione di altre tecniche di compressione. Molto usato è l'algoritmo chiamato "*Differential Pulse Code Modulation o DPCM*)" e suoi derivati.

- **LZW**. Questa codifica prende il nome dai suoi sviluppatori, A. Lempel, J. Ziv e Terry A. Welch. È un metodo di compressione di tipo generale e viene usata nei files GIF, TIFF e da programmi di utilità generale come *compress*.

Lossy Compression

Un procedimento di compressione si dice *lossy* o irreversibile quando parte dell'informazione viene deliberatamente scartata perché definita irrilevante, non percepibile, o tale da non compromettere in modo grave la sua percezione. In questo modo l'informazione originale verrà ricostruita con una certa approssimazione. Con questi sistemi si possono ottenere dei rapporti di compressione molto elevati (fino 100:1).
Nelle tecniche di tipo *lossy*, molto usate nelle aree dell'audio e del video digitale, si cerca di raggiungere lo scopo di massimizzare i benefici (elevato rapporto di compressione) al minor costo possibile (la perdita di qualità) tenendo conto di come verrà percepita l'informazione compressa dall'utilizzatore. Si parla, infatti, di codifica di tipo percettivo (***perceptual coding system***) e di modelli *psico-visivi* e *psico-uditivi*. Tra queste tecniche ricordiamo:

- **Subsampling:** Ovvero si ignorano certi dati. Come accennato in precedenza questo è un modo efficace di ridurre i dati, di solito applicato al colore, ed è attuato con l'uso dei formati 4:2:2 e 4:2:0 o altri ad essi analoghi. Questi non compromettono in modo rilevante l'informazione dato che la nostra percezione delle differenze di colore non è molto precisa. Il subsampling permette di ottenere una compressione dell'ordine di grandezza di 2:1 praticamente senza nessuna perdita rilevabile.

- **Quantizzazione Vettoriale (VQ):** Codifica nella quale insiemi di pixel, visti come vettori, vengono rimpiazzati da un singolo codice, il quale è un indice che punta ad una tabella di altri vettori, ognuno dei quali è un'approssimazione dell'insieme dato. Naturalmente lo stesso indice può essere usato per rappresentare più di un vettore.

- **Lossy Predictive Coding:** Questa tecnica è una variante della **DPCM**, alla quale è associata un certo livello di quantizzazione che introduce la perdita di informazione.

- **Discrete Cosine Transform (DCT):** Negli attuali sistemi di codifica è sicuramente la tecnica più usata Questa modalità non genera compressione, è il seguente stadio di quantizzazione, sempre ad essa associata, che la fa annoverare tra i sistemi con perdita di informazione. È in questa fase di quantizzazione che si tiene conto del modello percettivo del sistema visivo umano.

- **Subband Coding (SBC):** Con questa tecnica, molto usata in campo audio, una sorgente viene sottoposta ad una serie di operazioni (con un certo numero filtri) e suddivisa in bande con proprietà specifiche. Queste possono essere compresse con più efficienza dato che possiedono un livello di informazione minore della sorgente.

- **Discrete Wavelet Trasnsform (DWT):** La teoria relativa alle ***wavelet*** risale alla metà degli anni 80 e da allora si è evoluta come strumento matematico impiegato in varie aree di studio oltre a quello della compressione video. In pratica è un tipo di trasformata simile a quella di Fourier, che viene usata per approssimare una

funzione, ma invece di usare una base ortonormale infinita (seni e coseni) ne usa una finita.

- **Motion Estimation Techniques:** Queste tecniche si occupano dello sfruttamento delle ridondanze temporali presenti nelle sequenze video tenendo presente le caratteristiche percettive del sistema visivo umano.

Con nessuno delle singole tecniche citate si ottiene una riduzione di dati significativa. Il successo dei vari standard di compressione risiede nel fatto che essi combinano più tecniche in cascata, ed è questo che genera una significativa compressione.

Sistemi Simmetrici ed Asimmetrici

Gli algoritmi di compressione sono sistemi complessi e dispendiosi in termini calcolo. Quindi la scelta del livello di simmetria o asimmetria è importante.
Un sistema **simmetrico** è, per esempio, quello della videoconferenza dove la banda passante è limitata e le varie stazioni devono essere in grado di comprimere e decomprimere in un tempo breve le immagini e la voce. In altre parole la complessità della procedura di compressione è equamente divisa tra il codificatore ed il decodificatore. Nel caso dei sistemi **asimmetrici** vi è un solo codificatore e molti decodificatori (DVD, TV satellitare). Si ottiene un beneficio, riduzione del bitrate, semplificazione del decodificatore ed in genere una qualità migliore, nel rendere più semplice la procedura di decodifica a spese di un codificatore più complesso.

CAPITOLO 3

COMPRESSIONE SENZA PERDITA DI INFORMAZIONI
(Lossless Compression)

La necessità dei metodi di compressione lossless derivano dal fatto che in molte applicazioni non si vuole che tali algoritmi alterino i dati. Si pensi ad immagini tipiche del settore medico, all'uso del Fax o alla compressione dei testi. Diversi standard sono stati sviluppati per questi scopi, tutti hanno una base comune: La **Teoria dell'informazione**.

CODIFICA ENTROPICA

Nella teoria dell'informazione la quantità di informazione trasmessa da una sorgente E, (in bits) rappresentata da un insieme di simboli (alfabeto), è legato alla probabilità di questi simboli dalla seguente relazione:

$$I(E) = \log_2 \frac{1}{p(E)}$$

$I(E)$ definisce il livello di imprevedibilità dell'evento E (o auto-informazione di E). Un evento completamente prevedibile significa $p(E) = 1$ che vuol dire $I(E) = 0$ cioè nessuna informazione viene trasmessa.
Nel caso si volesse codificare un messaggio di una sorgente M, formato dai simboli $m_1\ m_2\ \ldots\ m_k$, aventi probabilità $p(m_k)$ si definisce **entropia** (Shannon) di M in bits/simbolo

$$H(M) = \sum_k p(m_k) I(m_k) = p(m_k) \log_2 \frac{1}{p(m_k)} = -p(m_k) \log_2 p(m_k);$$

cioè $H(M)$ è la media dell'informazione comunicata da ogni simbolo m_k.

Si abbia, ad esempio, una sorgente di informazioni M formato da un alfabeto di $N = 4$ simboli $\{a, b, c, d\}$ i quali sono equiprobabili. Quindi

$$p(a) = p(b) = p(c) = p(d) = \frac{1}{4} = 0,25$$

$$I(E) = \log_2 \left(\frac{1}{0,25} \right) = \log_2 (4) = 2$$

l'entropia risulta essere

$$H(M) = 0,25(2) + 0,25(2) + 0,25(2) + 0,25(2) = 2 \ \ bits / simbolo .$$

L'entropia in questo caso è massima ed il risultato coincide con $\log_2 N$ ossia è pari al numero di bit necessario per contare gli elementi dell'insieme. Nel caso si avesse invece una distribuzione di probabilità diversa come la seguente:

E	*p*(*E*)	1/*p*(*E*)	*I*(*E*)
a	0.6	1.667	0,737
b	0.2	5	2,321
c	0.1	10	3,322
d	0.1	10	3,322

Allora

$$H(M) = 0,6(0.737) + 0,2(2,321) + 0,1(3,322) + 0,1(3,322) = 1,570$$

Tale risultato è minore di $\log_2 N$ e quindi, suggerisce l'esistenza di un codice più efficiente per la nostra sorgente. Questo scopo viene raggiunto se codifichiamo i simboli con un numero di bits diverso, come il seguente

simbolo	**codice**
a	0
b	10
c	110
d	111

Il numero medio di bits per simbolo è definito come

$$L_{Medio} = \sum_i l_i p_i$$

dove l_i è la lunghezza del codice in bits relativa al simbolo $i-esimo$ e p_i la corrispondente probabilità. Nel nostro esempio la lunghezza media di questo codice è

$$L(M) = 0.6(1) + 0.2(2) + 0.1(3) + 0.1(3) = 1,6 \ \ bits / simbolo$$

il quale si avvicina molto al valore teorico dato dall'entropia. Questi risultati ci portano ad affermare che il numero medio di bits per simbolo usato per codificare un messaggio deve essere almeno uguale all'entropia della sorgente, $H(M) \le L(M)$; in altre parole per ogni sorgente di informazione esiste un limite inferiore alla comprimibilità, il valore della sua entropia H.

Codifica Huffman

Nel 1952 Huffman dimostrò come costruire un efficiente codice a lunghezza variabile (**VLC**) il quale però risulta valido solo se la distribuzione di probabilità della sorgente è conosciuta. In una sua variante l'algoritmo esamina la sorgente in due passi: Nel primo viene determinata la distribuzione statistica e nel secondo si effettua la codifica vera e propria.

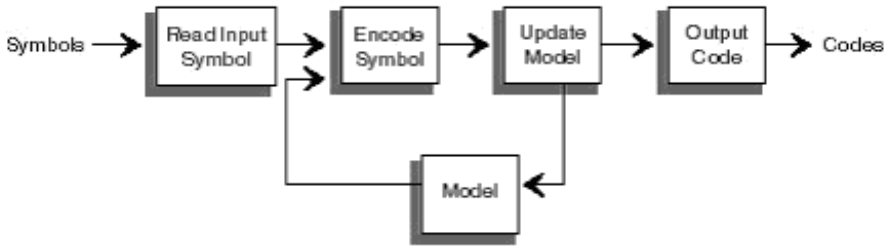


Figura 7: Schema di codifica dell'Adaptive Huffman Encoding

Un'ulteriore variante che non ha bisogno della conoscenza preventiva della distribuzione statistica è quella detta ***Adaptive Huffman Encoding*** nella quale il modello statistico di base viene continuamente aggiornato mentre l'algoritmo procede alla codifica dei simboli (Figura 7). Il decodificatore aggiornerà lo stesso modello di base alla stessa maniera tenendo conto dei codici già decodificati.

Codifica Aritmetica

Questa è una versione più sofisticata dell'approccio statistico di Huffman. In questo schema di codifica **intere sequenze** di simboli sono rappresentate da **singoli codici** secondo la loro probabilità. L'esempio della Figura 8 illustra l'idea base di questo sistema.
Si assuma un alfabeto $\{a,b,c\}$ con una distribuzione probabilistica come la seguente:

$$D = \{p(a) = 0,2,\ \ p(b) = 0,3,\ \ p(c) = 0,5\}$$

Si prenda, quindi, l'intervallo di numeri reali da $[0,1]$ e lo si suddivida secondo tali probabilità.

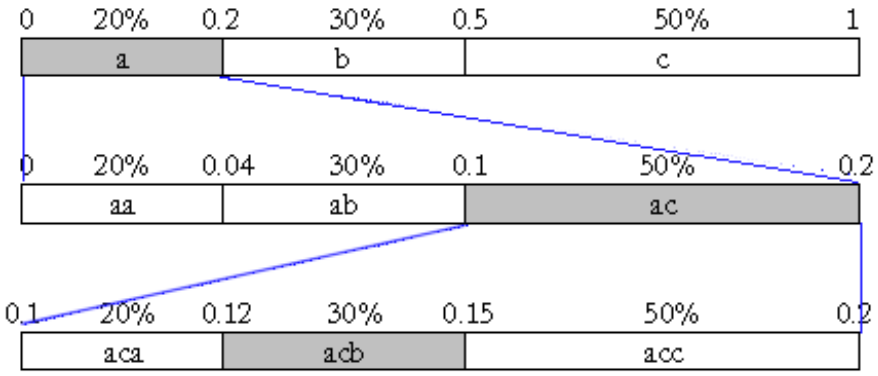


Figura 8: Esempio di codifica aritmetica

Per calcolare il codice, per esempio della sequenza acb l'intervallo di a viene suddiviso di nuovo secondo la distribuzione D, l'intervallo da 0,1 a 0,2 rappresenta la sequenza ac. Questo intervallo a sua volta è diviso secondo la medesima distribuzione D ottenendo per la sequenza acb l'intervallo da 0,12 a 0,15. Quindi qualsiasi numero in questo intervallo è il codice che rappresenta acb.
Questo codice è più efficiente di quello Huffman perché la sua caratteristica è quello di raggruppare i simboli dell'alfabeto.

CODIFICA PREDITTIVA

Finora abbiamo trattato sorgenti di informazione statisticamente indipendenti, cioè il valore di un dato simbolo non dipende dal valore di quelli precedenti. Molte sorgenti sono di tipo diverso; per esempio nella lingua italiana dopo la lettera q è quasi sicuro trovare la u, e che dopo qu è più probabile trovare una vocale di una consonante. Una sorgente dove la probabilità di un determinato simbolo dipende dai valori dei simboli precedenti è chiamata sorgente Markov. Se la dipendenza è relativa solo al valore precedente si parla di sorgente Markov del primo ordine, se la correlazione si estende fino ai due valori precedenti, si parla di sorgente del secondo ordine, e così via.
Le immagini fotografiche possono essere trattate come sorgenti Markov poiché è molto elevata la probabilità che l'intensità (la luminosità, il colore) di un pixel sia molto simile a quelli che lo circondano. La Figura 9 mostra l'elevata correlazione esistente tra l'intensità luminosa esistente tra pixel vicini.

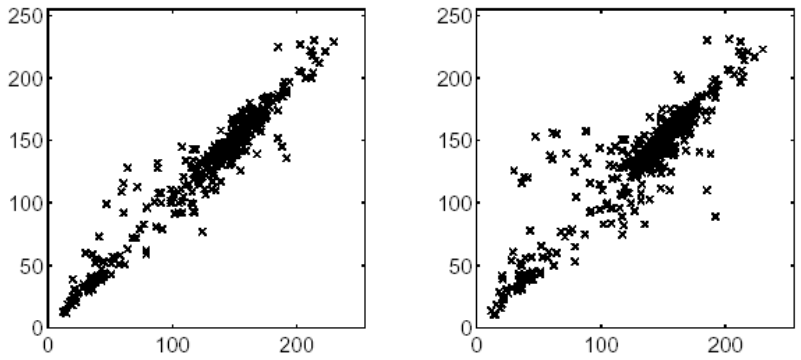


Figura 9: Correlazione tra pixel adiacenti a distanza 1 e 2

Il principio della codifica predittiva è semplice (Figura 11), invece di codificare la luminosità di un pixel facciamo una previsione sul suo valore in base a quelli dei vicini; si calcola la differenza, che chiameremo errore o residuo, tra il valore previsto e quello reale e assegniamo questo come valore del pixel dato. Per ricostruire l'immagine il decodificatore farà la medesima previsione che insieme all'errore ricevuto ricreerà la corretta luminosità dei pixel.

Di solito i valori associati ai pixel (secondo gli attuali standard) cadono nell'intervallo da 0 a 255. L'errore commesso dalla previsione cadrà nell'intervallo da –255 a +255. Ma come abbiamo già accennato data la natura dell'immagine la previsione sarà abbastanza accurata e quindi gli errori saranno molto prossimi a 0 (Figura 10). Abbiamo così ottenuto un insieme di valori con una distribuzione statistica non uniforme utile per essere sottoposta con efficienza ad una compressione di tipo entropico (codifica Huffman).
Questa tecnica predittiva è chiamata **Differential Pulse Code Modulation** il quale scopo, come abbiamo visto è quello di *decorrelare* i dati diminuendone l'entropia e quindi il contenuto di informazione.

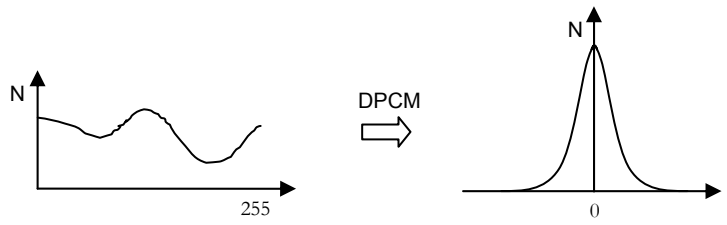


Figura 10: Applicazione dell'algoritmo DCPM ad un'immagine (N =numero di pixel aventi la medesima intensità)

Come algoritmo di previsione di un dato pixel si possono usare, come abbiamo accennato prima, funzioni (chiamate **predictor**) che coivolgono la conoscenza di più pixel adiacenti, è inoltre possibile usare più *predictor* in dipendenza dai valori assunti da questi pixel (**Adaptive Prediction**).

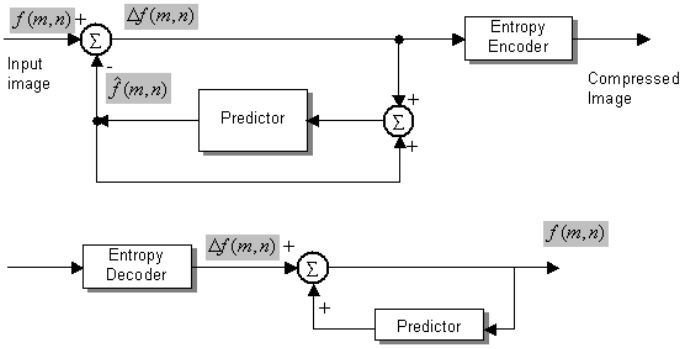


Figura 11: Principio della codifica DPCM (in alto) con il decodificatore (in basso)

RUN LENGTH ENCODING

Come menzionato in precedenza questo tipo di compressione rimpiazza valori ripetuti con una coppia <*lunghezza, valore*> dove per *lunghezza* è da intendersi il numero di volte che il *valore* è ripetuto. Una diretta applicazione di questo schema alle immagini non produce una soddisfacente compressione. Di solito il suo utilizzo riguarda contesti binari ove si possono trovare un numero rilevante di sequenze di "0" o "1". Per esempio: Lo standard bianco e nero per i Fax di Gruppo 3 e di Gruppo 4 o le immagini a scala di grigi visti come immagini *bit-plane*.

LZW

Questo è un sistema di compressione di tipo generale che fa uso di un dizionario il quale viene costruito "al volo" mentre vengono codificati i dati. L'idea è quella di non ripetere due volte la stessa sequenza di simboli. Ad ogni nuova sequenza viene associata una voce nel dizionario e l'indice che punta a quest'ultima è usato per codificare ogni altra eventuale identica sequenza futura. La codifica LZW è sfruttata in molti contesti data la sua efficienza e velocità.

CAPITOLO 4

COMPRESSIONE CON PERDITA DI INFORMAZIONE (Lossy Compression)

Tutte le tecniche di compressione con perdita di informazione operano in varie maniere una **quantizzazione**, cioè un determinato insieme di n valori (di solito di natura continua) viene rappresentato con un altro insieme di m valori dove $m < n$. Questa è la fase più importante in tutte le metodologie di compressione di tipo *lossy* poiché tale operazione deve essere compiuta tenendo conto di un modello dell'apparato psico-visivo (o psico-uditivo) dell'uomo per ottenere un prodotto la cui percezione sia la più possibile vicina a quella naturale.
Attualmente le tecniche di compressione di maggior successo utilizzano trasformate matematiche che hanno la caratteristica di far passare dal dominio del tempo a quello delle frequenze. Tra queste la trasformata di Fourier è sicuramente la più conosciuta ed è utilizzata come potente mezzo di analisi e sintesi di segnali e funzioni. Nel settore della compressione video è largamente usata una sua variante la *Discrete Cosine Transform* o **DCT**. Un'altra trasformata usata per scopi simili è la *Discrete Wavelet Transform* o **DWT**.

VALUTAZIONE DEGLI ERRORI

Per valutare la qualità di una tecnica di compressione di tipo *lossy* si può fare riferimento ad un criterio soggettivo oppure ad uno oggettivo.
Un criterio soggettivo coinvolge numerose persone le quali dovrebbero essere poste in una condizione visiva standard di illuminazione, distanza e tipo di display per dare un giudizio della *qualità percepita* secondo una pre-determinata scala (che potrebbe essere: eccellente, buona, accettabile, scadente, inaccettabile), naturalmente questa è una via onerosa e difficile da realizzare.
Di solito si usano dei criteri oggettivi. I più usati sono: La radice dell'errore quadratico medio $(RMSE)$, il rapporto segnale rumore di picco $(PSNR)$ ed il rapporto segnale rumore. Essi sono così definiti:

$$RMSE = \left[\frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (P_c(x,y) - P(x,y))^2 \right]^{\frac{1}{2}}$$

$$PSNR_{dB} = 10 \log \frac{\left[valore\ \ di\ \ picco \right]^2}{RMSE}$$

$$SNR_{dB} = 10 \log \left(\frac{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \left[P(x,y) \right]^2}{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (P_c(x,y) - P(x,y))^2} \right)$$

Dove $P(x, y)$ rappresenta l'immagine originale, vista come una matrice di pixels, ed $P_c(x, y)$ la stessa immagine prima compressa e poi decodificata.

È bene sottolineare che affidarsi a questo tipo di valutazioni sulla qualità delle immagini ha i suoi svantaggi. Da una parte semplici movimenti di rotazione, traslazione e variazioni di luminosità generano, con questi criteri, elevati errori ma non sono percepiti dal sistema visivo umano come una degradazione di qualità. Dall'altra parte piccole differenze tra i pixel possono apparire sgradevoli all'occhio umano. Un esempio sono gli artefatti di compressione come il *blocking* il quale è generato da una leggera differenza della luminosità tra blocchi di pixel adiacenti ma come risultato collaterale creano dei bordi che l'occhio è molto efficiente nell'evidenziare.

IL SISTEMA VISIVO (Human Visual System)

La destinazione finale di un'immagine statica o di una sequenza video è l'apparato visivo dell'uomo. Una comprensione delle sue caratteristiche e delle sue limitazioni è fondamentale nel progettare un sistema di compressione. L'informazione che deve raggiungere l'occhio è qualcosa che deve essere nel domino percettivo dello spettatore.

L'occhio

Questo è un organo molto complesso non a caso una buona parte del cervello è dedicata alla visione. Esemplificando si può dire che la parte sensibile alla luce, la retina[7], riceve i segnali luminosi che convertiti dai fotorecettori in messaggi nervosi giungono al cervello. Vi sono due famiglie di fotorecettori nella retina, i **coni** ed i **bastoncelli**. I coni sono deputati alla visione diurna ed al colore, sono circa 6-7 milioni e sono concentrati nella parte centrale della retina chiamata fovea, la quale si trova nelle immediate vicinanze dell'asse ottico. La loro concentrazione decresce velocemente fuori da questa area. I bastoncelli variano da 75 a 150 milioni, sono distribuiti più uniformemente sulla retina rispetto ai coni, sono scarsi nella fovea, hanno un massimo a circa 17°-20° rispetto quest'ultima e sono adibiti alla visione notturna (Figura 12).

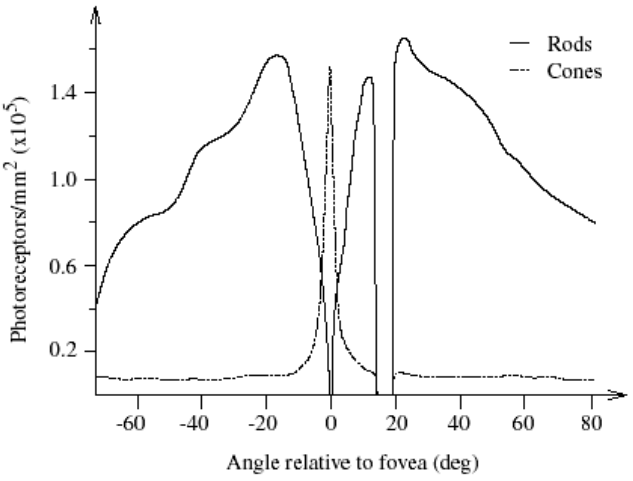


Figura 12: Distribuzione orizzontale dei fotorecettori

[7] La retina ha una struttura complessa stratificata. La luce prima di arrivare ai fotorecettori attraversa diversi tipi di cellule nervose tra loro connesse alcune dele quali inteconnettono sia i coni che i bastoncelli.

I bastoncelli procurano una visione monocromatica con una bassa risoluzione ma sono molto sensibili ed efficienti a luminosità molto basse; possono rilevare fino a 5 fotoni/sec (a 500nm); a questi livelli la fovea è praticamente cieca.
La visione diurna a colori è ad alta risoluzione ed è ristretta nello spettro tra 400 e 700 nm, in tale intervallo la sensibilità non è uniforme, esiste un picco nella regione del verde. Praticamente noi usiamo quasi esclusivamente la fovea per le nostre normali attività. I coni sono di tre tipi, ognuno dei quali ha un massimo di sensibilità nel Blu, nel Verde e nel Rosso. La luce di una sorgente provoca una risposta differenziata dei tre sensori che ci permette di dare un colore alla sensazione che suscitano (Figura 13).
Tutte queste informazioni, che fanno parte della *visione di basso livello*, sono raccolte dalla retina, codificate e convogliate per mezzo del nervo ottico verso la corteccia visiva del cervello ove hanno luogo i processi di *visione di alto livello*. Il rapporto di circa 100:1 tra il numero dei fotorecettori nella retina ed il numero di fibre del nervo ottico implica che una forma di compressione viene attuata in questo stadio. Si presume che questa efficiente compressione venga realizzata decomponendo l'immagine in un certo numero di canali percettivi (*Multi-Channel Decomposition*) i quali sono "sintonizzati" alle diverse caratteristiche spaziali, temporali e cromatiche in essa contenute.

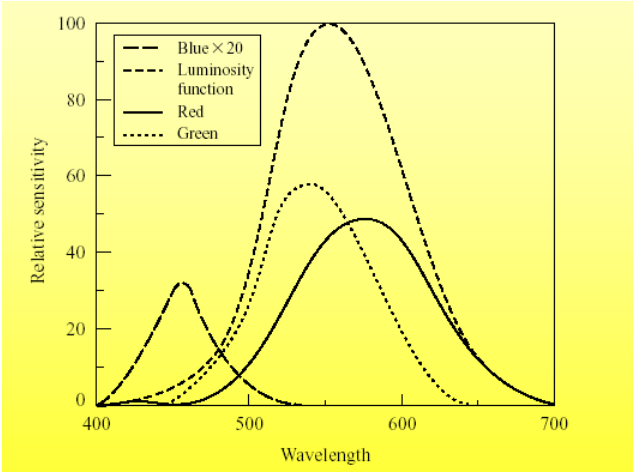


Figura 13: Sensibilità dell'occhio

Importanti caratteristiche della visione sono:

Acuità/Risoluzione. In pratica l'occhio è un apparato di campionamento spaziale con un passo[8] dato dalla distanza dei coni nella fovea Questa ha un'estensione limitata, copre circa 1-2°, per ottenere un campo visivo maggiore l'occhio si muove continuamente. Con i coni, è in grado di risolvere oggetti separati da circa 1 o 0.5 minuti d'arco. Tale risoluzione, in alcune particolari condizioni, può aumentare da 10 fino a 5 secondi d'arco (*Hyperacuity*) per mezzo di vari movimenti involontari di frequenza variabile (chiamati saccadici e micro-saccadici) che insieme all'effetto della persistenza della retina operano una sorta di sovra-campionamento.

Adattabilità. La retina è capace di adattare la propria sensibilità ai segnali che riceve con meccanismi che includono l'iride (apertura variabile) e i fotorecettori nella retina. Questo

[8] Supponendo i coni della stessa dimensione e a distanza costante nella fovea.

permette all'HVS di spaziare su un grande intervallo di intensità luminosa con un limitato numero di livelli.

Legge di Weber-Fechner. L'HVS può percepire piccolissime variazioni in luminanza tuttavia ciò dipende dal livello di luminosità di background. In altre parole, la percezione luminosa non è lineare ma logaritmica, siamo più sensibili alle variazioni nelle regioni più scure che in quelle più chiare. Questo implica che una stessa variazione in luminosità è più visibile nelle zone scure che in quelle chiare, questo effetto è chiamato *luminance masking*.

Curva di Sensibilità al contrasto. La Figura 14 mostra un'immagine di test sulla percezione visiva chiamata *funzione contrasto sensibilità* (CSF, Contrast Sensitivity Function), essa è costruita partendo dalla relazione

$$L = C\left(\frac{1}{2}\sin x\right)+\frac{1}{2} \qquad \begin{cases} 0 \le x \le 2\pi \\ 0 \le C \le 1 \end{cases} \qquad (1)$$

che definisce la luminosità L al variare di x e di C (fattore di contrasto).
L'immagine che ne deriva è un reticolo luminoso modulato con la funzione sinusoidale (1) a barre verticali chiare e scure che fornisce uno stimolo nel quale la frequenza spaziale varia in modo logaritmico da sinistra a destra e contemporaneamente il contrasto decresce, sempre in modo logaritmico, dal basso verso l'alto.
Per frequenza spaziale si intende il numero di cicli di barre verticali bianche e scure che sottendono un angolo di vista di un 1° (cpd, cycles per degree).

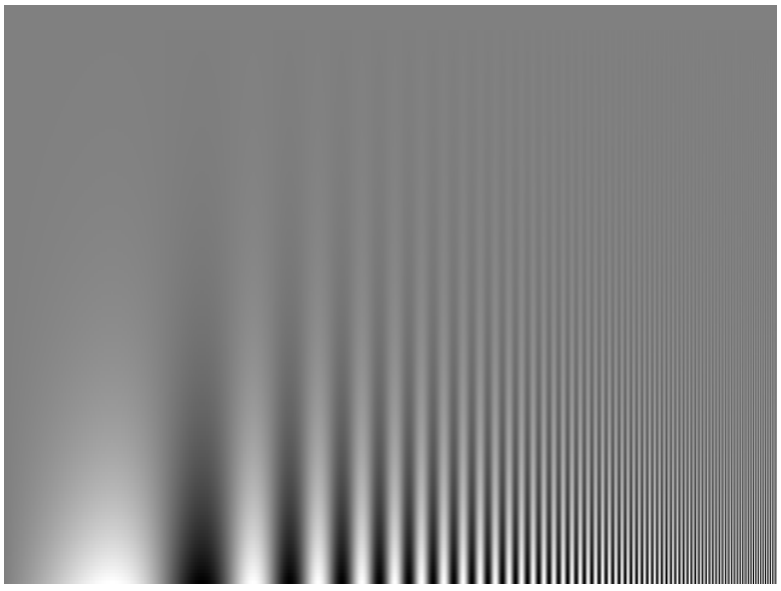


Figura 14: Test di sensibilità al contrasto di Campbell-Robson

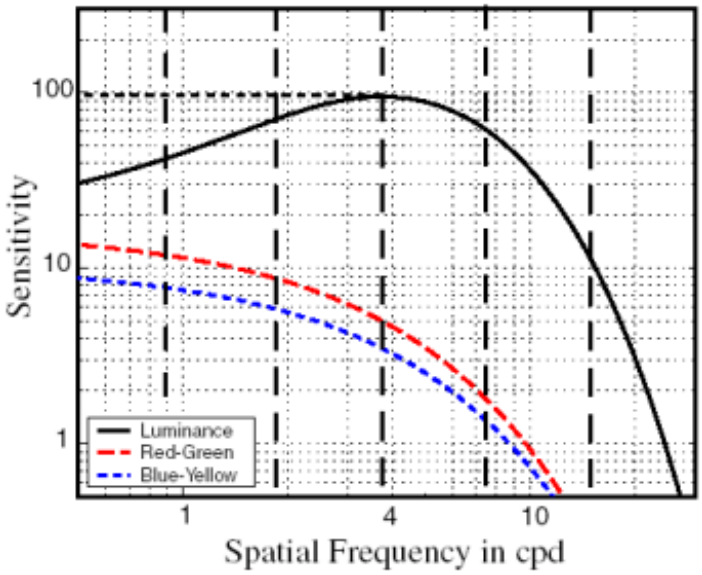


Figura 15: In ordinate la sensibilità (il reciproco del contrasto) ed in ascisse la frequenza in cicli/grado

La soglia di visibilità, definita dai punti oltre i quali le barre non sono distinguibili, descrivono una curva al di sopra della quale il nostro sistema visivo è completamente cieco, inoltre, la massima sensibilità viene raggiunta verso le frequenze di 5 cicli/grado e che il limite superiore per quanto riguarda la visibilità delle frequenze spaziali è di circa 60 cicli/grado. Nella Figura 15 sono anche riportate le curve relative ai colori ai quali l'apparato visivo mostra un'evidente minore sensibilità.
Inoltre anche l'orientamento del reticolo influenza la sensibilità, infatti, questa raggiunge un minimo quando le barre si trovano a 45°. Una conferma del fatto, accertato da tempo, che l'HVS è preferenzialmente "sintonizzato" su geometrie lineari di tipo verticale o orizzontale.

Proprietà temporali. Il sistema visivo mostra una risposta agli stimoli di tipo temporale molto simile a quella spaziale. Infatti, esiste l'analogo della CSF chiamata TSF (*Temporal Contrast Sensitivity Function*) dalla quale si deduce che la percezione di una sorgente luminosa variabile (*flickering light*) è funzione della sua frequenza e che esiste una frequenza detta *critical frequency flicker* (CFF) oltre quale la sorgente viene percepita come continua.
Questa è una diretta conseguenza dell'effetto di persistenza delle immagini sulla retina. La frequenza CFF non è costante ma aumenta, arrivando fino a 100 Hz, in funzione della luminosità ambientale.
Questo è un "difetto" molto utile che ci impedisce di "vedere" le luci, la TV o i display dei computer "sfarfallare" e di poter apprezzare un film al cinema.

Mascheramento. Questo tipo di effetto (*texture masking*) esprime il fatto che la percezione di un segnale A è in qualche modo inibito da un altro segnale detto di mascheramento B. Per gli scopi della compressione delle immagini questo è molto importante poiché vuol dire che in determinati casi gli errori di quantizzazione possono essere resi non percepibili.
Questo fenomeno risulta evidente se si prendono in considerazione le parti omogenee e le parti attive (presenza di dettagli) di un'immagine. Nelle parti uniformi (basse frequenze spaziali) è facile individuare delle distorsioni (errori) mentre in una regione attiva (elevate frequenze spaziali) queste possono essere facilmente nascoste.

CODIFICA PREDITTIVA (lossy)

Questa tecnica (Lossy Predictive Coding) è basata sulla codifica predittiva di cui abbiamo parlato in precedenza, la quale è, originariamente, di tipo senza perdita di informazione (*lossless*). Si tratta in pratica di aggiungere al sistema della codifica predittiva uno stadio di quantizzazione che permette di ottenere una compressione maggiore con una qualità ragionevole (Figura 16).

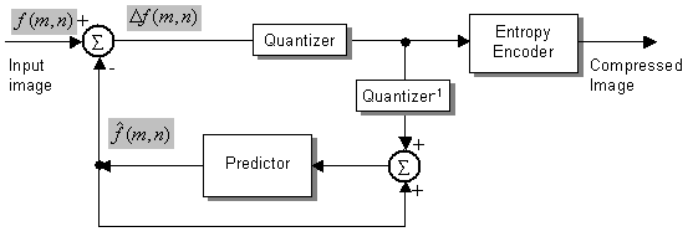


Figura 16: Codificatore DPCM con stadio di quantizzazione

Con questo sistema si ottengono ottimi risultati mediante tecniche di quantizzazione di tipo dinamico; applicando, cioè, una quantizzazione più fine nelle zone dell'immagine uniformi ove gli errori di codifica risultano più evidenti, ed una più ampia nelle aree più attive (elevate frequenze spaziali) ove l'effetto mascheramento rende gli errori meno percepibili.

DISCRETE COSINE TRANSFORM

Questa trasformata è una variante di quella più conosciuta dovuta a J. B. Fourier e che nel nostro caso ha il compito di mappare le intensità dei pixel nel formato YUV o YCbCr, dal dominio dell'immagine bidimensionale a quello delle frequenze spaziali. In analogia alla mappatura tra il dominio del tempo e quello delle frequenze nel caso dello sviluppo di Fourier.

Esistono ben collaudati algoritmi che applicano questa procedura, sono anche disponibili sistemi hardware dedicati a questo scopo. L'impiego di questa metodologia è relativamente semplice, come si può vedere nello schema a blocchi nella Figura 17. Il procedimento è il seguente:

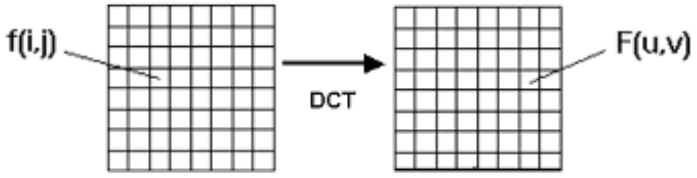
1. **Scomposizione dell'immagine a blocchi** – L'immagine, composta da campioni di luminanza e crominanza, è suddivisa in blocchi elementari di 8×8 pixel. Questa dimensione si è rivelata ottimale per questo tipo di applicazione.
   Nel campionamento 4:2:2 ci sono 6480 blocchi valori di luminanza (Y) e 3240 blocchi di campioni per ogni componente del colore (U, V o Cb, Cr). Ognuno dei 64 numeri variano da 0 a 255 per Y e da –128 a +127 per quelli relativi al colore U e V.

2. **DCT** - Ad ognuno di questi blocchi di 64 valori viene applicata la trasformata DCT:

$$F(u,v) = \frac{1}{4} C_u C_v \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cos\left(\frac{(2x+1)\pi u}{16}\right) \cos\left(\frac{(2y+1)\pi v}{16}\right)$$

$$C_u,\ C_v = \begin{cases} \frac{1}{2}, & if\ \ u = v = 0 \\ \frac{1}{\sqrt{2}}, & if\ \ u = 0\ \ o\ \ v = 0 \\ 1, & if\ \ u \neq 0\ \ e\ \ v \neq 0 \end{cases}$$

Si ottengono altri 64 valori che rappresentano i coefficienti delle frequenze spaziali tranne il primo $F(0,0)$ chiamato DC che riproduce la media dei valori delle intensità luminosa del blocco.



3. **Quantizzazione** – Questo passo tiene conto delle caratteristiche del sistema visivo dell'uomo (HVS), in particolare il sistema occhio-cervello non distingue i dettagli minuti (elevate frequenze spaziali) di un'immagine al disotto di una certa luminosità. Per sfruttare questa caratteristica percettiva ogni coefficiente $F(u,v)$ può essere diviso per un valore fisso globale (*quantizzazione scalare uniforme*); oppure ognuno dei coefficienti è diviso per i valori di una tabella di fattori predeterminata (*quantizzazione scalare non uniforme*); infine vengono posti a zero tutti i coefficienti inferiori ad un certo valore di soglia predefinito.
   Il valore del primo coefficiente DC viene codificato con molta precisione con l'algoritmo DCPM in relazione al blocco precedente. Questo riduce gli artefatti (*blocking artifacts*) quando l'immagine viene ricostruita.

4. **Zig-zag scanning** – A differenza del primo coefficiente del blocco (DC), gli altri, chiamati coefficienti AC, sono serializzati nel modo indicato nella figura e trasformati in un vettore di 63 elementi. Questo per renderli idonei alla successiva codifica entropica (RLE /VLC) e per sfruttare il fatto che il vettore conterrà sequenze di zeri.

5. **RLE** e **VLC –** Questi due fasi finali sono di solito accoppiati in un unico algoritmo con l'ausilio di una tabella di conversione.

I coefficienti $F(u,v)$ rappresentano i fattori di un insieme di funzioni di base i quali possono essere visti come insiemi di valori dei pixel. Nel nostro caso abbiamo 64 matrici relative a questi valori. L'operazione inversa DCT consiste nel moltiplicare ogni coefficiente per la rispettiva funzione di base e sommare tutte e 64 insiemi di valori dei pixel.

$$f(x,y)=\frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7}C_uC_vF(u,v)\cos\left(\frac{(2x+1)\pi u}{16}\right)\cos\left(\frac{(2y+1)\pi v}{16}\right)$$

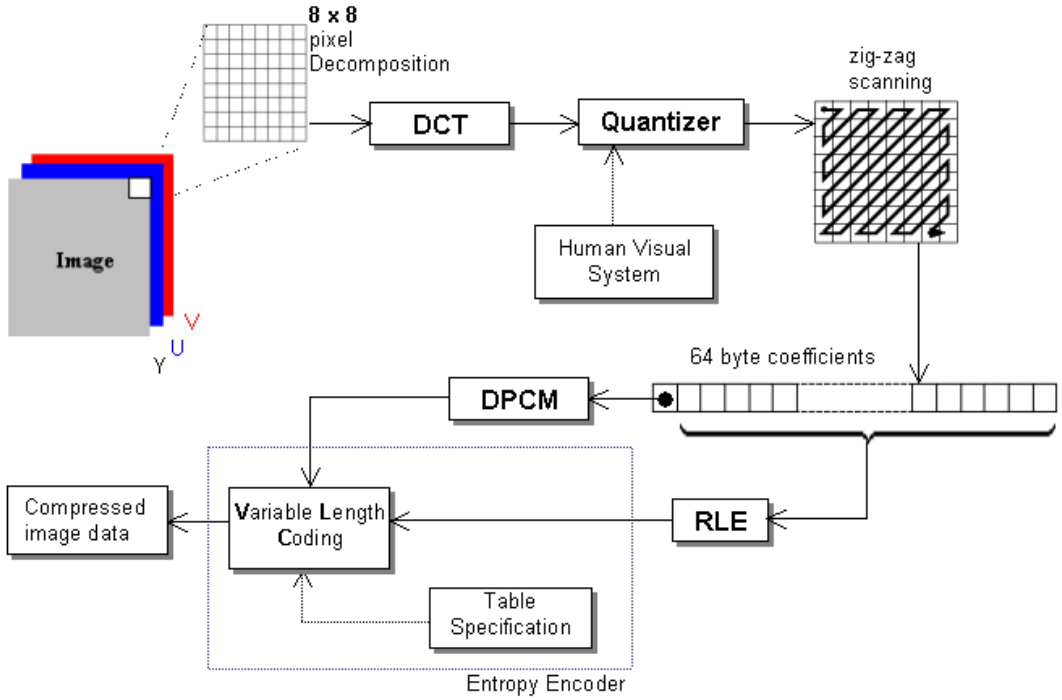


Figura 17: Algoritmo di compressione basato sulla DCT.

QUANTIZZAZIONE VETTORIALE

A differenza della quantizzazione **scalare** che opera sul valore dell'intensità del singolo pixel, quella vettoriale (**Vector Quantization** o **VQ**) prende contemporaneamente in considerazione i valori di gruppi di pixel $m \times n$ considerati come vettori. Sono definiti dei vettori di ricostruzione rappresentati da un indice che formano una tabella di riferimento. La quantizzazione è attuata confrontando un vettore di input con ognuno dei vettori di ricostruzione e scegliendo quello che minimizza l'errore quadratico medio od un altro criterio di qualità analogo.

La quantizzazione vettoriale è una tecnica molto potente e produce risultati molto prossimi ai valori limite teorici ma risulta molto onerosa dal punto di vista computazionale. La parte realtiva alla generazione e gestione della tabella dei vettori di ricostruzione può essere molto complessa.

Esiste un esempio di questa tecnica molto conosciuto e riguarda il formato dei file GIF (Gaphics Interchange Format) sviluppato da CompuServe.

Un'immagine RGB (24 bits, 16 milioni di colori) è rappresentata nel formato GIF da un insieme di indici a 8 bits che puntano ad una tabella di 256 colori. L'algoritmo deve trovare un insieme di 256 colori che meglio approssima, con una distorsione minima, l'immagine data.

CODIFICA A SOTTOBANDE (SBC)

Questa tecnica, molta usata in campo audio, trae vantaggio dal fatto che i segnali non hanno una distribuzione spettrale uniforme. Nella forma più semplice questa codifica consiste nel decomporre il segnale in un certo numero di bande le quali sono trattate in modo indipendente. Ci saranno bande con una quantità di energia elevata, altre nelle quali ce ne sarà di meno ed altre ancora ove sarà nulla. Come risultato la codifica di ogni banda darà un contributo differente, in termini di bits, alla codifica totale ottenendo una compressione più efficiente.
La scelta del numero di sottobande dipende dai successivi stadi compressione che verranno usati e dal tipo di informazione che essi rappresentano.
L'operazione di decomposizione in sottobande viene attuata con l'ausilio di filtri digitali che operano su dati discreti. Dato un filtro la cui risposta è rappresentata da un insieme di valori H_P esso viene moltiplicato con tutti gli N valori x_n i quali rappresentano l'andamento di un dato segnale campionato ad una certa frequenza *f*, questa operazione detta di *convoluzione* genera un output di N valori y_n.

$$y_n = H * x = \sum_{k=0}^{P-1} H_k x_{n-k}$$

La risposta del filtro H_P può essere di tipo passa-basso o passa-alto, questo, insieme alla loro applicazione iterativa al segnale permette di suddividere le sue componenti spettrali in bande di ampiezza sempre più piccole.

Questo tipo di codifica è importante per la stretta relazione con la trasformata **DWT**. In pratica alcune forme di decomposizione con le wavelet sono realizzate appunto con banchi di filtri digitali.

DISCRETE WAVELET TRASFORM o DWT

La trasformata di Fourier o FT è la più famosa tra le trasformate lineari, essa permette di passare da una rappresentazione nel piano delle ampiezze-tempo di un segnale o di una funzione, ad un'altra nel piano ampiezza-frequenza. L'inconveniente di questa trasformazione è la perdita dell'informazione relativa alla posizione (temporale o spaziale) corrispondente ad una data frequenza.
Per ovviare a questa limitazione è stata introdotta la Short-Time FT (STFT) la quale opera una trasformazione del segnale per parti. In pratica viene definita una funzione "finestra" $g(\tau - t)$ di dimensione opportuna la quale scorre lungo il segnale; la trasformata è applicata di volta in volta al segmento di segnale entro l'estensione di tale finestra. Con questa nuova variante della FT si evidenzia il contenuto spettrale all'interno della finestra temporale Δt nell'intorno del tempo t. Tuttavia anche questo approccio ha dei difetti:

- Fissata la funzione finestra sono parimenti determinate le risoluzioni o le indeterminazioni su Δt nel dominio del tempo e Δf in quello delle frequenze; ne consegue che due eventi distanti meno di Δt o Δf nei rispettivi domini non possono essere discriminati;

- Le due indeterminazioni sono inversamente proporzionali, una delle due risoluzioni può essere migliorata ma solo a scapito dell'altra e viceversa;

- Essendo le due risoluzioni fisse i rapporti $\Delta t / t$ e $\Delta f / f$ risultano variabili.

Quest'ultimo punto è cruciale una cosa è identificare una frequenza di 100MHz con una precisione di $\Delta f = 1Khz$ un'altra identificare una frequenza di 100KHz con la medesima precisione.
Il problema è risolvibile apportando un'ulteriore modifica alla STFT in modo da poter usare delle funzioni di finestra che possano non solo scorrere lungo il segnale ma anche avere un'estensione temporale (larghezza) variabile. Questa è l'idea base che ha portato all'introduzione della trasformata wavelet.
A differenza di quella di Fourier questa nuova trasformata adotta una funzione di finestra prototipo chiamata la *mother wavelet*. Traslando questa funzione si ottengono informazioni di tipo temporale; dilatandola o contraendola si estrae il contenuto spettrale.
Con una scelta opportuna di questa funzione wavelet si può far in modo di far variare in modo dinamico la risoluzione nel piano tempo-frequenza mantenendo costante il rapporto $\Delta f / f$ ottenendo così un'analisi del segnale con differenti risoluzioni. In questo modo, al diminuire della frequenza l'individuazione del contenuto spettrale si fa più preciso, mentre la precisione nella localizzazione temporale aumenta alle frequenze più alte.

Per quanto riguarda le applicazioni nel campo della compressione digitale questo significa che le alte frequenze, corrispondenti a transienti audio o bordi ed aree attive nelle immagini, vengono trasformate con wavelet brevi con le quali si ottiene una buona localizzazione temporale o spaziale. Le frequenze basse sono invece trasformate con wavelet lunghe le quali danno una migliore risoluzione in frequenza.
L'interesse nell'applicare questo tipo di trasformata in questo campo è duplice. Prima di tutto possiede un elevato potere di impacchettamento dell'energia, e quindi dell'informazione; ne consegue che i coefficienti della trasformata diversi da zero saranno relativamente pochi, essendo la maggior parte molto prossimi a zero. In più, della stessa immagine si possono ottenere versioni con risoluzioni differenti.

L'implementazione della DWT, nel campo della compressione digitale, è attuata con le tecniche, conosciute da tempo, di *Subband Coding* utilizzando dei ben definiti filtri digitali di tipo FIR (*Finite Impulse Response*) chiamati QMF, dall'inglese *Quadrature Mirror Filter*. Il processo è di tipo iterativo e consiste nel far passare ripetutamente un segnale attraverso dei filtri passa-alto e passa-basso.
Nella Figura 18 è illustrata un esempio unidimensionale di questo tipo di tecnica. Durante la prima iterazione il vettore rappresentativo del segnale (valori campionati) è fatto passare attraverso una coppia di filtri i quali dividono la banda passante originaria in due sottobande rappresentate da due vettori di dimensione uguale a quello originario contenenti le componenti a bassa frequenza (L) e quelle ad alta frequenza (H). Tali bande, a causa dell'azione dei filtri, sono dimezzate, quindi risultano sovracampionate. Questo permette, in base al teorema di Nyquist, di scartare la metà dei campioni. Il vettore dei componenti di alta frequenza è chiamato vettore dei dettagli ed i suoi valori sono i coefficienti della trasformata wavelet. La seconda e successive iterazioni sono eseguite sul vettore dei componenti di bassa frequenza delle iterazioni precedenti.

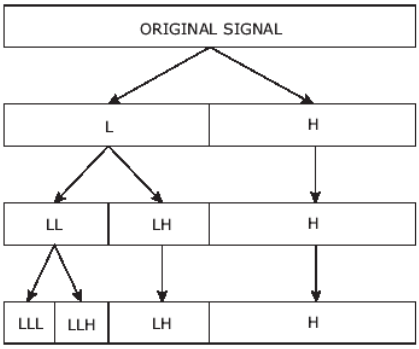


Figura 18

Nella Figura 19 è mostrata la decomposizione di un'immagine bidimensionale. Il risultato della prima iterazione produce quattro immagini con una risoluzione di ¼ rispetto all'originale. Si noti come l'immagine contenente le basse frequenze, che sarà l'input dello stadio successivo, è ancora facilmente riconoscibile. Le altre tre immagini rappresentano i dettagli (le alte frequenze) nelle tre orientazioni verticali, orizzontali e diagonali.

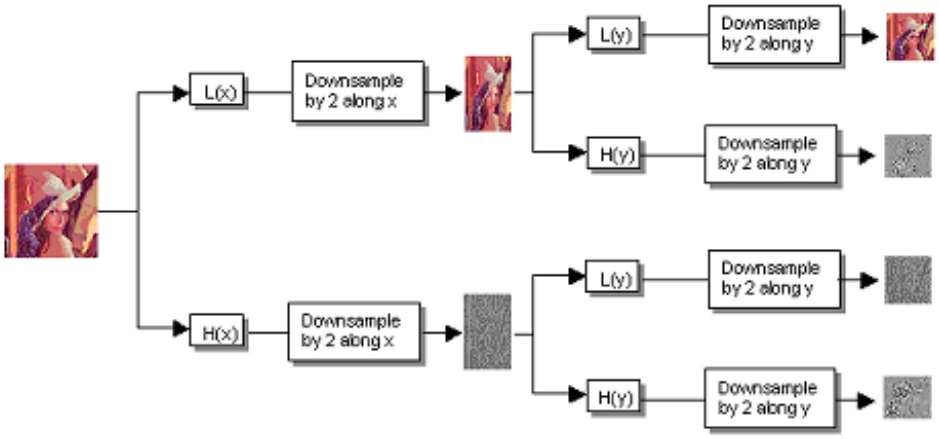


Figura 19

Come esempio più completo si osservi il risultato dell'applicazione di tre stadi di questa procedura all'immagine di "Lena" di Figura 20.

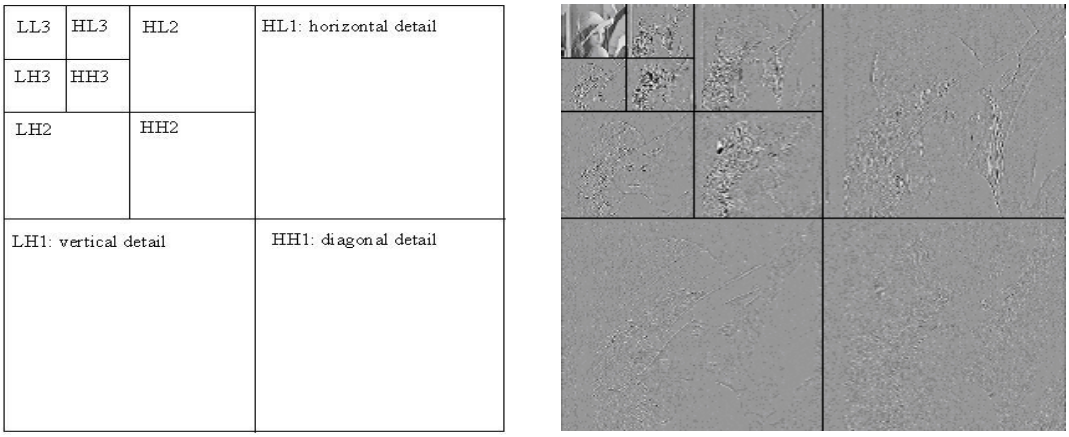


Figura 20

Durante il primo passaggio i filtri dividono l'immagine nelle componenti HH1, HL1 e LH1 di alta frequenza ed una L1 di bassa sequenza; la loro risoluzione è ridotta di $1/4$ rispetto a quella iniziale. Questi componenti non saranno più analizzati mentre la componente a bassa frequenza viene nuovamente sottoposta al processo di filtraggio. Questo crea un'altra serie di immagini di alta frequenza HH2, HL2 e LH2, ed una di bassa frequenza L2 tutte ridotte ad $1/16$. L2 è di nuovo filtrata ottenendo altre componenti, ridotte ad un $1/64$, di alta frequenza HH3, HL3 e LH3 ed una in bassa frequenza LL3.
Si notino gli effetti di questo tipo di decomposizione sull'immagine iniziale. Si ottengono immagini con risoluzione decrescente (sempre più piccole) i cui coefficienti (valori relativi ai pixel) sono diversi da zero in quelle parti dell'immagine corrispondenti ai bordi o ai dettagli. Tali immagini mostrano una statistica non uniforme che li rende degli ottimi candidati ad un'efficiente compressione di tipo entropico. Anche la rimanente immagine di bassa frequenza è molto facile da comprimere date le dimensioni molto ridotte.
In fase di decodifica si procede con le stesse operazioni a ritroso. Le immagini HH, HL, LH e LL formano l'input di un banco di filtri, questa volta di sintesi, i quali insieme alle operazioni di sovracampionamento ricreano l'immagine originaria.

In base alle caratteristiche delle wavelet già citate Shapiro progettò nel 1993 un algoritmo di compressione basato su questa trasformata chiamato *Embedded Zerotree Wavelet* o **EZW**. Tale metodo tiene conto delle seguenti osservazioni:

- Più alto è il valore numerico di un coefficiente wavelet, più rilevanza possiede e quindi più informazione contiene. Questo implica che l'algoritmo di codifica deve dare una priorità maggiore a tali coefficienti;

- L'energia di un'immagine si concentra prevalentemente su pochi coefficienti. Nella Figura 20 la maggioranza dei coefficienti è nullo;

- Esiste una forte similitudine tra tutte le sottobande della medesima orientazione. I valori massimi e medi dei coefficienti tendono a diventare più piccoli muovendosi dalle frequenze più basse, fattore di scala elevato, verso le più alte, fattore di scala minore. Questo implica che se un coefficiente non è significativo ad una determinata scala, allora è molto probabile che lo saranno anche tutti quei coefficienti che si trovano nella stessa posizione ma a scale minori.

L'output prodotto dall'algoritmo EZW è di tipo progressivo; questo significa che più dati vengono aggiunti al processo di compressione maggiore sarà il dettaglio nel ricostruire l'immagine. O detto in altra maniera in fase di decompressione si può scegliere la risoluzione finale dell'immagine in base alle proprie esigenze.

MOTION ESTIMATION

Nella maggior parte delle volte una sequenza video è formata da una successione di immagini (fotogrammi o frame) molto simili tra loro, le differenze sono dovute in genere a traslazioni nel loro contenuto. Queste ridondanze temporali sono sfruttate con le tecniche di stima dei movimenti (**Motion Estimation)**, le quali calcolano gli spostamenti (**Motion Vectors**) che si sono verificati nei contenuti di una scena, tra un frame ed il successivo in modo da minimizzare la loro differenza (**Motion Compensation**). Questo aspetto della compressione video è una delle parti più onerose in termini di calcolo.
In pratica il frame corrente è scomposto in blocchi di pixels e preso in considerazione uno di questi, il problema consiste nel trovare un blocco identico o molto simile a quello dato nel frame precedente, detto frame di riferimento. Si attua, in pratica, una sorta di predizione sfruttando l'informazione del frame di riferimento.
Per valutare se due blocchi sono simili si ricorre a dei criteri oggettivi di tipo numerico i quali calcolano la Differenza Minima Assoluta (MAD) o il Minimo Errore Quadratico Medio (MMSE) tra i due blocchi presi in considerazione. Lo scopo è quello di trovare il vettore spostamento (u,v) che rende minima una di queste quantità.

$$MAD(u,v) = \frac{1}{mn} \sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} \left| C(i,j) - R(i+u, j+v) \right|$$

$$MMSE(u,v) = \frac{1}{mn} \sum_{i=-m/2}^{m/2} \sum_{-n/2}^{n/2} [C(i,j) - R(i+u, j+v)]^2$$

dove

$C(i,j)$ blocco di coordinate i,j (angolo in alto a destra) nel frame corrente
$R(x,y)$ è un blocco nel frame di riferimento
(u,v) sono i *motion vectors*
$m \times n$ è l'area di variabilità

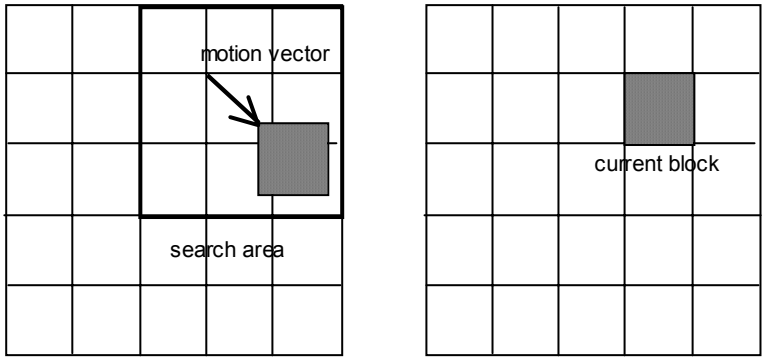
Il metodo più semplice è quello detto **block matching** nel quale il frame corrente è di solito scomposto in blocchi di 16×16 pixels, chiamati **macroblock**. Per ognuno di questi si deve localizzare, nel frame di riferimento, un blocco identico o simile, entro una zona limitata chiamata area di ricerca. Una volta individuato un possibile candidato si possono presentare tre casi:

1. Il contenuto dei due macroblocchi è identico (**close match**): il macroblocco del frame corrente viene sostituito dal solo vettore spostamento rispetto al frame di riferimento.

2. Il contenuto dei due macroblock è molto simile (**best match**): al macroblock relativo al frame corrente viene sostituito il vettore spostamento e la differenza o residuo (**prediction error**) tra i due macroblock.

3. Il contenuto dei due macroblock è molto diverso: il macroblock viene codificato normalmente.

Nei primi due casi si ottiene una notevole diminuzione di informazione e quindi una maggiore compressione.
La parte decodificatrice per la ricostruzione usa i vettori spostamento per recuperare i macroblock dal frame di riferimento ai quali aggiunge l'eventuale residuo per ottenere il macroblock corrente.



Esistono delle varianti di questo procedimento con lo scopo di diminuire la quantità di calcoli necessari ed ottenere nel contempo buoni risultati. Esempi di questi algoritmi sono il block matching gerarchico (*Hierarchical Block Matching*) e la ricerca logaritmica bi-dimensionale (*2-D Logarithmic Search*).

Oltre alle tecniche di *block matching* ne vengono usate altre, un esempio sono il *Gradient Matching* e la **Phase correlation** (correlazione di fase).
Il metodo della correlazione di fase è molto efficiente e sfrutta la proprietà della trasformata di Fourier di far corrispondere spostamenti spaziali a spostamenti di fase nel dominio delle frequenze. Si procede come segue:

1. Si scelgono due insiemi di pixels di uguale dimensione e posizione appartenenti a due frame o a due semiquadri (fields) adiacenti. Ai valori di questi due insiemi si applica la trasformata di Fourier;

2. Si calcola la differenza tra queste due trasformate e si converte tutto in coordinate polari (ampiezza e fase);

3. Si opera la normalizzazione delle ampiezze, mentre per ogni frequenza componente le fasi di una delle trasformate sono sottratte da quelle dell'altra;

4. Con il risultato si calcola l'anti-trasformata di Fourier con la quale si ottiene un nuovo insieme di valori di pixels.

Questi nuovi valori associati ai pixels formano una superficie detta di correlazione di fase (*correlation surface*) la quale presenta dei picchi che corrispondono alla direzione ed all'ampiezza degli spostamenti relativi degli oggetti presenti negli insiemi dati. In pratica rappresentano dei *possibili* vettori spostamento (*candidate vectors*).
Questo, però, non dà nessuna informazione sulla posizione assoluta dove tali spostamenti, all'interno dell'insieme, sono avvenuti.
Per ottenere la posizione degli oggetti interessati dallo spostamento uno degli insiemi viene traslato rispetto all'altro in una direzione e con un'ampiezza corrispondente ai picchi della

superficie di correlazione fino ad ottenere una correlazione tra di essi. Le correlazioni positive permettono di associare i possibili vettori alle posizioni degli oggetti che hanno manifestato quello spostamento.

Con i metodi citati le previsioni sugli spostamenti avviene a livello di pixel. Per ottenere migliori risultati nella stima dei vettori spostamento, anche se a costo di una maggiore complessità, si ricorre a tecniche di interpolazione che permettono di determinare tali vettori con un'accuratezza di *frazioni di pixel* (mezzo pixel, un quarto di pixel e perfino un ottavo di pixel).

CAPITOLO 5

COMPRESSIONE AUDIO

La compressione audio come tecnica è più vecchia di quella video ed è stata largamente applicata in campo analogico, col termine "*riduzione del rumore*", nei vari sistemi audio Dolby A, B e C, chiamati anche compressori di dinamica (audio companding).
La rappresentazione digitale del suono consiste in una serie valori campionati ad intervalli regolari (frequenza di campionamento). Di solito le frequenze di campionamento usate variano da 8000 Hz, con una qualità bassa di tipo telefonico, fino a 48000 o 96000 Hz con una qualità elevata tipica dell'audio cinematografico. L'accuratezza con la quale l'audio viene riprodotto dipende anche dal numero di valori (discreti) che ogni campione può avere, generalmente il numero di bits varia da 8 (256 valori) o 16 bits (65536 valori).

Il sistema Uditivo

In analogia con il sistema visivo anche per il suono è fondamentale conoscere il comportamento del sistema uditivo in modo da sfruttarne le caratteristiche. Il sistema uditivo umano è capace di notevoli performance. Ha un campo di percezione, nelle regioni più sensibili, di oltre 100 dB con un campo di frequenza da 20 Hz a 20.000Hz (10 ottave). Il potere di discriminazione della combinazione orecchio-cervello ha dell'incredibile, possiamo distinguere una conversazione o sentire pronunciare il nostro nome in ambienti con elevati rumori di fondo ed altre interferenze ad un livello tale da sembrare una sfida alle leggi della Fisica. Tutto questo presuppone un sofisticato sistema di analisi ed elaborazione della quale si possono fare solo delle ragionevoli ipotesi.

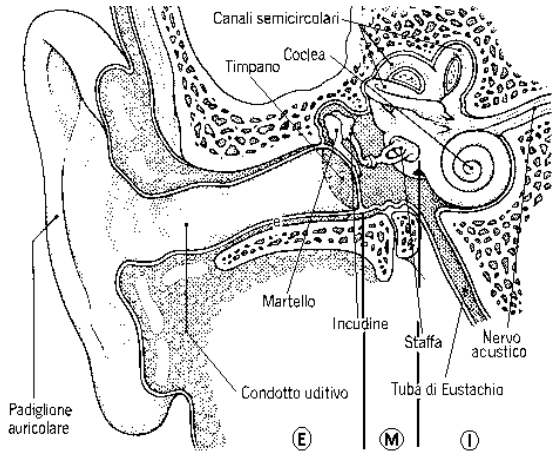


Figura 21: Orecchio e le sue parti

L'orecchio, l'organo dell'udito si veda la Figura 21, è normalmente suddiviso in tre parti. L'orecchio Esterno (E) formato dal *padiglione auricolare*, dal *condotto uditivo* e dal *timpano*. L'orecchio medio (M) che comprende gli "ossicini": *martello*, *incudine* e *staffa*. L'orecchio Interno (I) dove sono presenti la *coclea*, *i canali semicircolari* ed il *nervo acustico*.
Il suo funzionamento, in linee generali, è il seguente: le onde sonore (onde di pressione) convogliate dal padiglione auricolare e amplificate dal condotto uditivo (dalla forma a tromba) mettono in vibrazione il timpano il quale è collegato al sistema degli *ossicini* nell'orecchio medio. Questi formano un complesso di leve che amplificano e adattano il segnale in modo che sia percepibile dall'orecchio interno (in pratica avviene un adattamento di impedenza). La staffa (orecchio medio) trasferisce le vibrazioni alla coclea (orecchio interno) tramite una membrana: la *finestra ovale*. La coclea è l'organo sensoriale vero e proprio ove le onde sonore vengono trasformate in segnali nervosi che il nervo acustico trasporta al cervello.

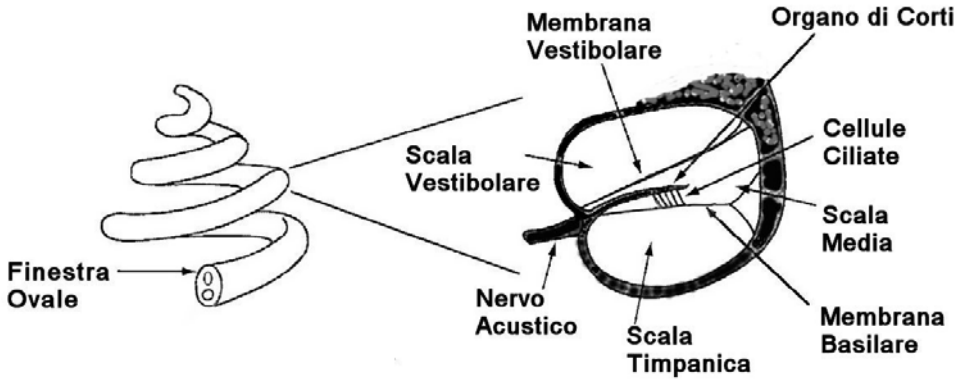


Figura 22: La coclea è una cavità a forma chiocciola pieno di liquido. A destra nella figura è schematizzata una sezione trasversale.

La coclea, schematizzata nella Figura 22, è un canale a forma di chiocciola divisa in senso longitudinale dalla *membrana basilare* e dalla *membrana vestibolare* in tre cavità contenenti liquido, chiamate *scala vestibolare*, *scala media* e *scala timpanica*. La parte sensibile detta *Organo di Corti* è costituito da una serie di *cellule ciliate* disposte lungo la membrana basilare. Le cellule ciliate rappresentano i recettori acustici che trasformano il suono in impulsi nervosi. La vibrazione acustica dalla finestra ovale si propaga nel liquido contenuto nella cavità cocleare e quindi alla membrana basilare che mette in movimento le cellule cigliate le quali generano l'impulso nervoso. In pratica la membrana basilare, di larghezza e spessore variabile, funge da analizzatore di spettro, le sue parti risuonano a frequenze diverse. Alle frequenze alte (20KHz) vicino alla finestra ovale e alle frequenze basse (20 Hz) verso l'apice della chiocciola.

Il Modello Psico-Acustico

La sensibilità dell'udito ai suoni non è lineare ma dipende dalla frequenza e dalla loro intensità. La Figura 23 mostra il diagramma di Fletcher e Munson ove sono riportate le curve di risposta di uguale percezione sonora (isofoniche) al variare della frequenza. La curva più bassa rappresenta la soglia di udibilità quella più alta quella del dolore.
Nel grafico si può notare che le frequenze basse (minori di 200-250Hz) devono essere di intensità superiore rispetto a quelle alte per fornire una sensazione uguale.
Nell'Intervallo tra 1 e 4 KHz c'è una zona di risposta costante che coincide con le frequenze del linguaggio parlato.
Tutti i sistemi di compressione devono tenere conto di questo fatto, inoltre, un'altra caratteristica dell'udito del quale si deve tenere conto è l'effetto di mascheramento, nella

compressione audio tale effetto è sfruttato in modo estensivo. Il mascheramento può essere in relazione alla frequenza od al tempo.

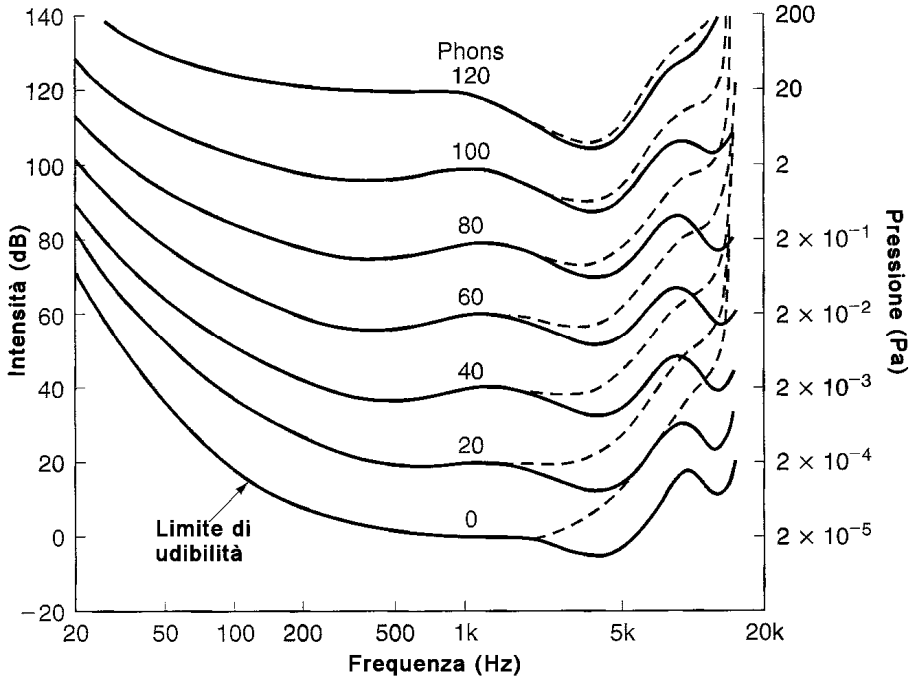


Figura 23: Diagramma di Fletcher e Munson con le curve di uguale percezione sonora

Il **mascheramento simultaneo** (**Frequency masking**) avviene quando sono presenti due suoni simultanei, vicini in frequenza ma di diversa intensità; in questo caso il suono corrispondente al segnale di intensità più elevata maschera quello vicino in frequenza diminuendone (parzialmente o totalmente) la percettibilità. Negli algoritmi di compressione il segnale mascherato può essere scartato o codificato con un numero minore di bits ottenendo una diminuzione dell'informazione.

Il **mascheramento temporale** (**Temporal masking**) consiste nel fatto che l'orecchio diviene insensibile per un certo tempo prima e per un certo tempo dopo uno stimolo sonoro intenso. In modo simile all'andamento di un suono che ha un tempo di *attacco*, aumento dell'intensità col tempo, ed un tempo di *decadimento*, diminuzione dell'intensità col tempo.
Il mascheramento dovuto al decadimento (**Forward masking**) può arrivare fino a 200ms mentre quello di attacco (**Backward masking**) è dell'ordine delle decine di ms. Naturalmente qualsiasi suono presente durante tali intervalli non verrà preso in considerazione dagli algoritmi di compressione ottenendo anche in questo modo una diminuzione dell'informazione.

Il mascheramento simultaneo dipende dalla particolare natura della membrana basilare la quale può essere suddivisa in zone dette *bande critiche* (**critical bands**) le quali entrano in risonanza a frequenze differenti. L'ampiezza di tali zone è funzione anch'essa della frequenza, esse diventano più ampie verso le alte frequenze. Se due suoni eccitano la medesima zona ed uno di questi è molto intenso l'altro risulta molto attenuato o non udibile (Mascheramento Simultaneo). Inoltre la membrana basilare, dopo essere stata eccitata, impiega del tempo per ritornare in equilibrio; questo rende conto dell'effetto del mascheramento temporale.

STANDARDS AUDIO

La quasi totalità degli algoritmi di compressione audio usa la tecnica della codifica a sottobande o sue varianti. Questa, infatti, meglio si adatta alle caratteristiche percettive dell'udito. Nella Figura 24 è schematizzato un tipico compressore audio.

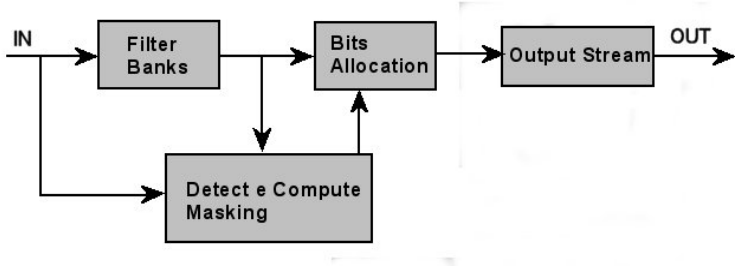


Figura 24: Schema di un tipico algoritmo di compressione Audio

MPEG-1

Lo standard audio MPEG-1 è un gruppo di schemi di compressione suddivisi in *layers:* I, II e III. Tali layers sono compatibili verso il basso, cioè un decodificatore di *layer* II è in grado di decodificare anche il *layer* I. L'intervallo di *bitrate* disponibile va da 32 a 224 Kbit/s per canale mentre le frequenze di campionamento permesse sono 32000, 44100 e 48000 Hz. L'audio può essere di tipo monofonico o stereo nelle seguenti varianti:

- Monofonico per canale audio a sorgente unica.
- Due canali audio indipendenti. Simile alla modalità stereo ma usato per esempio per due lingue diverse
- Il modo stereo classico, il quale vuol dire fornire un effetto tridimensionale all'apparato uditivo umano.
- Lo Joint stereo che trae vantaggio dalla caratteristica che hanno i suoni di alta e bassa frequenza di non possedere localizzazione spaziale. Nella loro codifica, quindi, tali frequenze vengono associate ad uno solo dei due canali.
- Il modo stereo M/S (Mid/Sideband) il quale sfrutta le ridondanze presenti nei due canali audio. Infatti, invece di codificare i canali L (sinistro) e R (destro) vengono codificate la somma L+R e la differenza L-R, quest'ultimo risulta essere un segnale piccolo e quindi codificabile in modo aggressivo.

In tutti i *layers* una grande importanza riveste il banco di filtri (Polyphase Filter Bank) usato per dividere il segnale audio in 32 sottobande di uguale larghezza. Questo è un buon compromesso con le caratteristiche dell'apparato uditivo umano. Infatti, come abbiamo già accennato, le "bande critiche" dell'apparato uditivo non sono di uguale larghezza ma dipendono dalla frequenza, ne deriva che a frequenze basse le sottobande comprendono più bande critiche.
Il primo layer (**Layer I**) è il più semplice e fornisce dei buoni risultati per *bitrate* di 192 Kbit/s (384 Kbit/s per lo stereo) con un rapporto di compressione di circa 1:4. E' stato usato dalla Philips nei nastri DCC (Digital Compact Cassette). L'algoritmo codifica in *frames* di 384 campioni audio raggruppati in 12 campioni da ogni sottobanda. Il principale vantaggio del layer I è la relativa semplicità di codifica e decodifica.
Il secondo (**Layer II**) è un pò più complesso di quello precedente ed è stato progettato per ottenere dei buoni risultati a bitrate di 128 Kbit/s per canale (rapporto di compressione di

circa 1:6) con un massimo di 192 Kbits/s. A differenza del layer I la codifica avviene con frames di 1152 campioni per canale audio, cioè 3 gruppi di 12 campioni per ognuno delle 32 sottobande. E' usato per la trasmissione dell'audio digitale e nei Video-CD.
Il **Layer III** è il sistema più complesso dal punto di vista algoritmico ed è anche il famoso formato chiamato universalmente **MP3**. Esso usa un diverso modello psicoacustico che cerca di compensare alcuni difetti dell'analisi con i filtri aggiungendo una successiva elaborazione per mezzo della trasformata DCT. In pratica viene applicata una MDCT (Modified Discrete Cosine Transform) alle uscite del banco di filtri allo scopo di ottenere una migliore risoluzione spettrale. Naturalmente questo layer offre una migliore qualità anche a *bitrate* bassi, intorno ai 64Kbits/s per canale (rapporto di compressione di circa 1:11).

MPEG-2

Lo standard MPEG-2 audio estende quello dell'MPEG-1 al fine di fornire il supporto per le seguenti estensioni:

- Sei canali audio (5+1 channel). Cinque canali audio alta fedeltà (centro, destro, sinistro, surround sinistro, surround destro) più un canale per le basse frequenze.
- Fino a sette canali audio per il supporto multilingue
- La possibilità di utilizzare frequenze di campionamento più basse (16000, 22050 e 24000 Hz.)
- Compatibilità verso il basso con il MPEG-1

Di recente si è aggiunto un altro standard l'MPEG-2 AAC (Advanced Audio Coding) che non è compatibile a ritroso con l'MPEG-1. Questo nuovo standard fornisce un audio multicanale di alta qualità anche a bassi *bitrate* (64 Kbits/s per canale).

Dollby AC3

La tecnologia Dolby Digital AC-3 multicanale della Dolby Laboratories usata nei film su DVD, nelle trasmissioni satellitari ed in altri campi, è una variante dei sistemi audio delle sale cinematografiche. Il marchio Dolby è, infatti, strettamente legato all'evoluzione della riproduzione audio in questo settore sin dagli anni 80. Attualmente lo standard audio analogico per i film a 35mm è il Dolby SR (Spectral Recording) il quale fornisce fino a quattro canali (Left, Center, Right, Surround ed uno opzionale per i bassi) codificati sulle due tracce ottiche della pellicola (Figura 25).

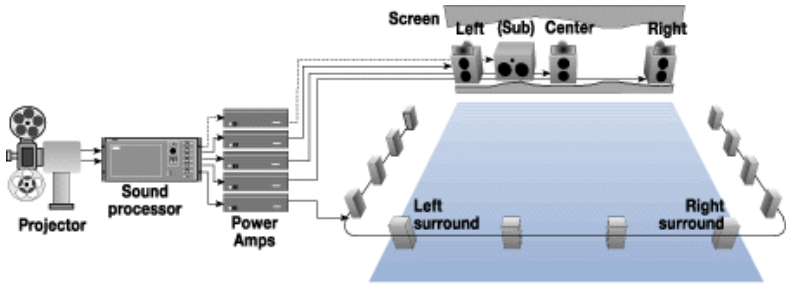


Figura 25: Sistema Dolby SR analogico per sale cinematografiche

Il canale denominato "Surround" si riferisce al posizionamento dei relativi diffusori "intorno" all'ascoltatore.
Nel 1992 la Dolby introdusse un sistema digitale più efficiente e completo con l'idea di affiancarlo a quello analogico. La cosa fu accolta con molto favore ed oggi è il sistema più utilizzato in tutto il mondo sia nelle sale cinematografiche che negli studi professionali fino al comune utilizzatore. I canali forniti dal Dolby AC-3 sono 6 ai quali si fa riferimento con la sigla 5.1: sinistro, centrale, destro, surround sinistro, surround destro e LFE. Quest'ultimo deriva dalle iniziali dei termini Low Frequency Effects e si riferisce alla riproduzione delle basse frequenze audio (<120Hz). Nella Figura 26 è schematizzato un tipico sistema Dolby per uso "casalingo".

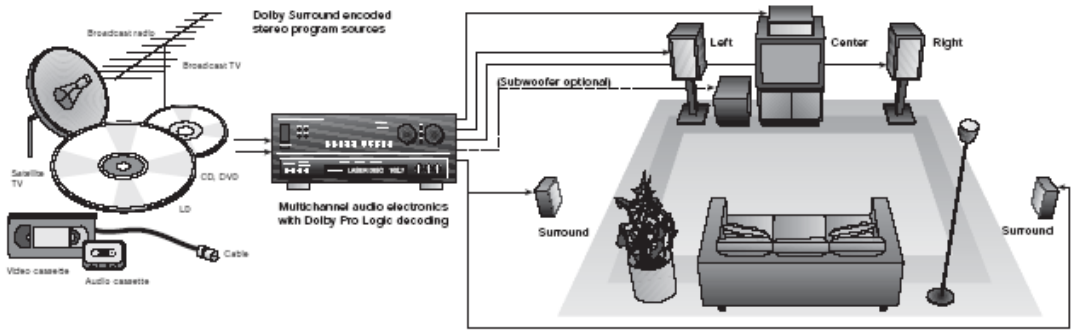


Figura 26

L'architettura del sistema Dolby ha come obbiettivo di limitare drasticamente la banda (max 384 o 448 Kbps, con campionamento a 48000Hz), per riprodurre vari canali audio senza alterare in modo significativo la qualità del segnale originario. Inoltre il codificatore AC-3 può essere visto come una soluzione audio completo e versatile dal quale è possibile estrarre il flusso audio che meglio si adatta alle nostre esigenze (downmixing stereo o mono).

DTS

Negli anni 90 fu introdotto un altro sistema multicanale per il cinema, il DTS (Digital Theater Surround) della Digital Theater Systems. Anche in questo caso fu derivata una variante per il mercato "home theater" chiamato *Coherent Acoustics System*. Dal punto di vista dell'ascoltatore è quasi indistinguibile dall'AC-3 ma è stato creato con un diverso obbiettivo, quello di offrire un audio multicanale di qualità entro i limiti di banda imposti dal CD-ROM[9]. Le tecniche di codifica sono mirate allo sfruttamento delle ridondanze (con l'uso estensivo delle tecniche ADPCM). Inoltre, la caratteristica di asimmetria tra codificatore e decodificatore sono state aumentate a scapito di un codificatore più complesso con il risultato di un aumento di versatilità di tutto il sistema.

[9] Questa scelta deriva dal fatto che la traccia audio digitale DTS delle sale cinematografiche non si trova sulla pellicola ma su CD-ROM. Nella pellicola è solo registrato un codice di sincronizzazione per tenere allineato l'audio con il video.

CAPITOLO 6

STANDARDS di COMPRESSIONE

IMMAGINI E VIDEO

JPEG

JPEG (Joint Photographic Experts Group) è una commissione unificata ISO (International Organization for Standardization) e ITU-R con il compito di stabilire uno standard per la compressione delle immagini statiche. Nel 1992 l'JPEG è stato riconosciuto come standard mondiale. L'implementazione corrente di questo standard ha molte opzioni e permette sia la compressione di tipo **lossless** (usando la codifica predittiva) che quella di tipo **lossy.** Quest'ultima è anche chiamata **baseline JPEG**. La tecnica utilizzata è quella della divisione dell'immagine in blocchi 8x8 pixels, l'applicazione della trasformata DCT seguita da una quantizzazione e dalla codifica entropica, come mostrato nella Figura 17.
Con l'opzione **Baseline Sequential Mode**, l'immagine è codificata e decodificata dall'angolo in alto a sinistra, da sinistra a destra, e dall'alto in basso; mentre con l'opzione **Progressive Mode** viene prima codificata e decodificata una versione dell'immagine a bassa qualità alla quale sono aggiunti i dettagli con passaggi successivi. Esistono, inoltre, estensioni di questo standard che ne migliorano le caratteristiche al costo di una diminuzione del rapporto di compressione.

Lo standard JPEG ha naturalmente dei difetti, quelli più gravi sono: l'algoritmo non è efficiente quando deve comprimere dei grafici o del testo; a compressioni elevate si notano degli artefatti come il **blocking** ed ha delle limitazioni nella gestione di immagini di grandi dimensioni. Per ovviare a queste debolezze è stato proposto un nuovo standard l'**JPEG2000**. Esso cerca di risolvere tutti i difetti del JPEG e introduce l'utilizzo della trasformata **Wavelet**. Alcuni suoi punti di forza sono:

- Compatibilità con la versione precedente dello standard JPEG;
- Un comportamento migliore per tassi di compressione elevati;
- Può gestire agevolmente immagini di grandi dimensioni;
- Su una stessa immagine è possibile combinare compressioni di tipo lossy e lossless;
- Una maggiore protezione contri gli errori;
- Codifica e Decodifica ROI (**Region Of Interest**) la quale permette di comprimere, a scelta, determinate parti dell'immagine con una qualità maggiore.

Inoltre l'utilizzo delle **DCW** elimina l'insorgere degli artefatti di blocking anche ad elevati rapporti di compressione.

H.261

Questa raccomandazione dell'ITU-T del 1990 fa riferimento alla parte della compressione video per le videoconferenze (e video telefonia) su linee ISDN. Il bitrate minimo, incluso l'audio, è di 64 Kb/s od un multiplo intero di tale velocità (fino a 30). In pratica per una videoconferenza di buona qualità bisogna arrivare ad almeno 384 Kb/s cioè tre link ISDN (da 64 Kb/s+64Kb/s). Il formato video è il **CIF** [10]di 352 x 288 pixels oppure il **QCIF** 176 x 144 pixels con un frame rate variabile da 7 a 30 fps. La raccomandazione H.261 è stata la base per la costruzione dello standard MPEG1.

MPEG-1 e MPEG-2

MPEG sono le iniziali del **M**oving **P**ictures **E**xperts **G**roup una commissione costituita dall'ISO per definire uno standard di compressione audio e video. L'approccio adottato da questo standard è innovativo, non viene definito come deve essere fatto un codificatore bensì stabilisce come deve essere strutturato lo *stream* dei dati ovvero come deve essere decodificato tale flusso (Figura 27). I vantaggi sono evidenti; gli algoritmi di compressione possono cambiare e migliorare ma finché il flusso aderisce alle specifiche, cioè è *MPEG compliant*, qualsiasi decoder potrà decodificarlo.

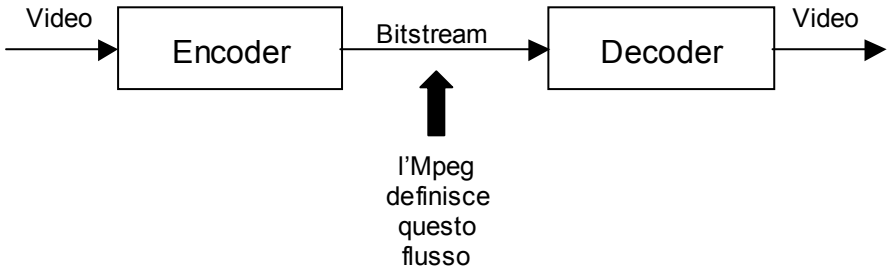


Figura 27: Cosa lo standard MPEG definisce.

Nello standard MPEG è definita una gerarchia di elementi come mostrato nella Figura 28. A partire dal livello più alto al più basso sono: *Sequence*, *Group of Pictures o GOP*; *Frame* o *Picture*, *slice*, *macroblock* e *block*.
Una *Sequence* è una successione di immagini di lunghezza arbitraria, un esempio è un video clip, un intero programma televisivo o un film.

[10] CIF sta per Common Intermediate Format. La necessità di tale formato si è resa utile per favorire l'interoperabilità a livello mondiale dei sistemi di videoconferenza data la presenza di due standard video quello a 625/50 (righe/semiquadri al secondo) e quello a 525/60.

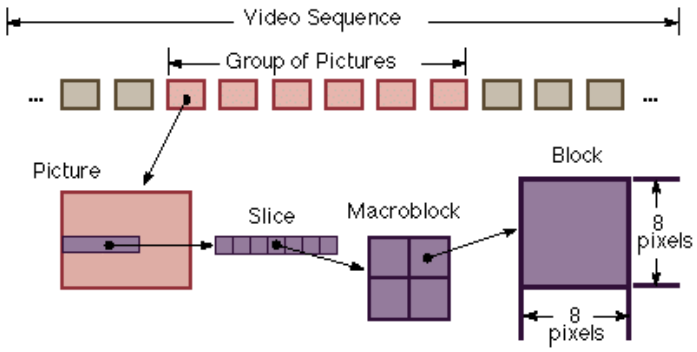


Figura 28: La struttura gerarchica dello standard MPEG.

Una *Sequence* è formata da un gruppo ripetitivo di frame chiamata *GOP*, questa a sua volta, è costituita da una serie di *frame* ed ogni frame da un certo numero di *slice*.
Uno *slice* è formato da un insieme di *macroblock* (16x16 pixels) ed un *macroblock* da quattro *block* (8x8 pixels).
Ogni *frame* può essere codificato in tre modi diversi che vengono così chiamati:

- Frame di tipo **I**. Questi frame detti anche **Intra-frame** sono codificati spazialmente (**Intra-coded**) senza nessun riferimento ad altri frame, sono il punto di riferimento di tutti gli altri tipi.

- Frame di tipo **P**. Questi frame chiamati anche **Predicted Frame** sono il frutto di una previsione basata su un **I-frame** o un altro **P-frame** precedente (**forward prediction**). Questi frame, insieme a quelli di tipo **B**, tengono conto delle ridondanze temporali utilizzando le tecniche di **motion estimation e compensation**.

- Frame di tipo **B**. Questi frame detti anche **Bidirectional Frame** sono costruiti tenendo conto sia dei frame **I** o **P** precedenti (**forward prediction**) che dai frame **I** o **P** che lo seguono (**backward prediction**) (Figura 29).

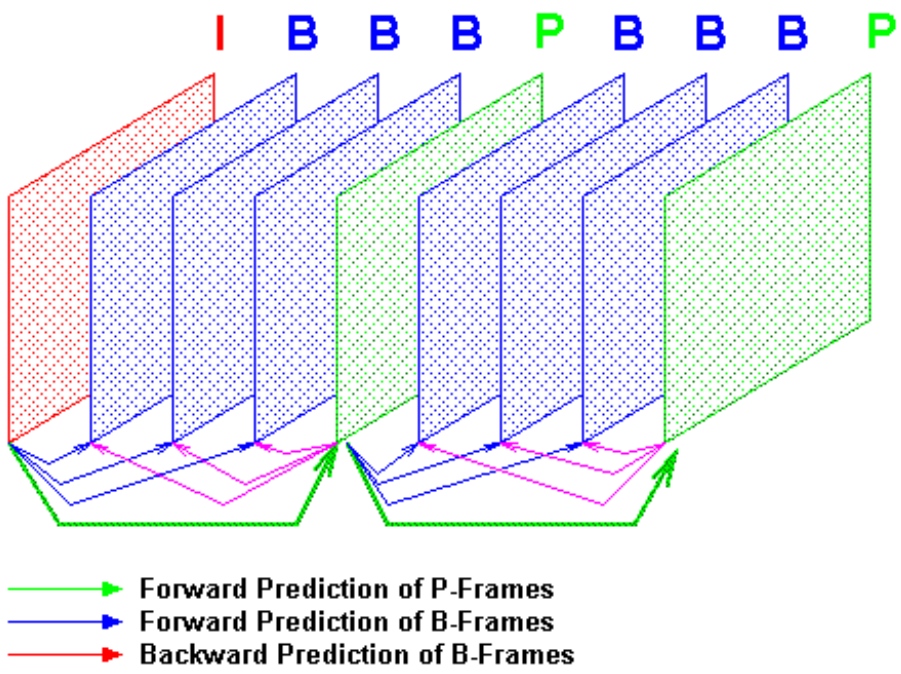


Figura 29: I tre tipi di frame nello standard MPEG

L'introduzione dei frame di tipo **B** tiene conto del fatto che parti del contenuto di un frame possono non essere disponibili nel frame precedente ma sono presenti nel frame successivo. Questo accade, per esempio, quando oggetti presenti in un fotogramma video si muovono evidenziando un *background* che prima non era visibile (Figura 30).

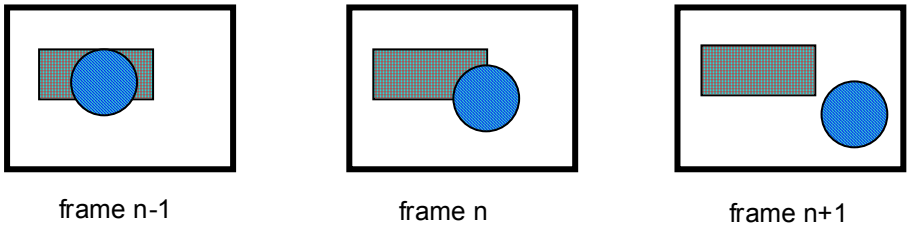


Figura 30: Utilità dei frame di tipo B. Nei tre frame consecutivi il cerchio muovendosi rivela parti del rettangolo che possono essere recuperati solo dal frame n+1.

Il concetto di **GOP** è molto importante in MPEG, esso consiste in una sequenza fissata di immagini costituita da frame **I**, **P** e **B** (per esempio: IBBPBBPBBPBB). Ogni GOP inizia con un frame **I**, i successivi sono predetti in base a quest'ultimo, e sono caratterizzati da due parametri M e N. M stabilisce la distanza tra i frame **I**, e N la distanza tra i frame **P**. Un GOP, inoltre, può essere chiuso od aperto.
Un GOP è chiuso, e quindi indipendente dagli altri, quando tutte le predizioni avvengono entro i limiti del medesimo GOP; in questo caso l'ultimo frame è di tipo P. È invece aperto quando le predizioni possono estendersi oltre il singolo GOP, questo significa che l'ultimo frame è di tipo B il quale dipende anche dall'I-frame del GOP seguente.
Si noti che la sequenza di visualizzazione delle immagini nel GOP è diversa da quella di codifica e decodifica. Infatti, data la dipendenza temporale dei frame B da quelli I e P questi

ultimi dovranno essere già disponibili nel decodificatore per poter ricostruire i frame B.
Nella Figura 31 è schematizzato un tipico decodificatore MPEG.

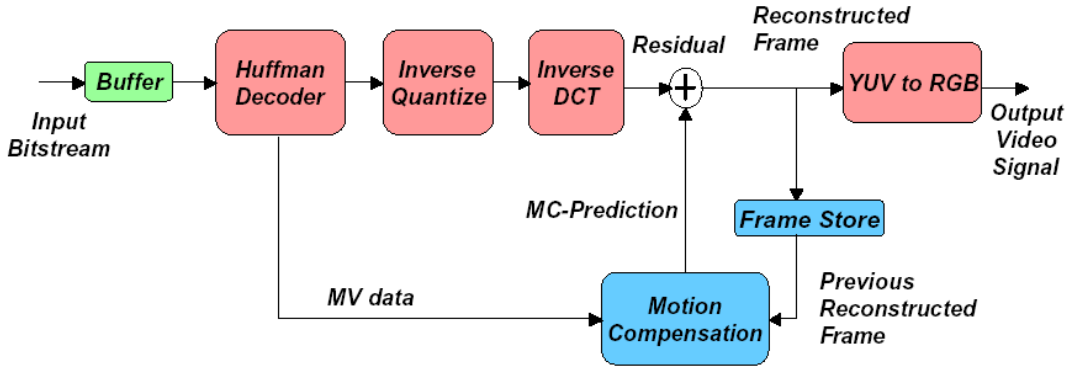


Figura 31: Decoder MPEG

Nel 1992 il **M**oving **P**ictures **E**xperts **G**roup ha rilasciato lo standard **MPEG-1** studiato principalmente per applicazioni video su CD-ROM e quindi con un bitrate massimo di 1,5 Mbit/s. Per ottenere questo risultato la risoluzione video è stata limitata al formato cosiddetto **SIF**[11] (*Source Input format*) di 352**X**288 pixels per i sistemi con frequenza di quadro di 25 Hz e di 352**X**240 per quelli a 30 Hz.
Lo standard è diviso in tre parti che definiscono, oltre il tipo di compressione video, quella audio ed il sistema di **multiplexing** cioè la tecnica di fusione dei due flussi audio e video (*elementary stream*) in un unico flusso (*Program Stream)* in modo da poter essere correttamente riprodotti contemporaneamente.
L'MPEG-1 è lo standard utilizzato nei VideoCD il quale fornisce una qualità video simile a quello delle cassette VHS.

Nel 1994 il **M**oving **P**ictures **E**xperts **G**roup ha rilasciato lo standard **MPEG-2** inizialmente pensato per le trasmissioni televisive con bitrate fino a 10 Mbit/s. Successivamente lo standard venne ampliato per utilizzarlo nel sistema TV ad alta definizione HDTV. L'MPEG2 è stato adottato dall'ITU-T per le telecomunicazioni con la raccomandazione **H.262**.
Lo schema di codifica dell'MPEG-2 è un'estensione di quello dell'MPEG-1 con alcuni miglioramenti ed una particolare attenzione alla gestione del video interlacciato e nella stima dei vettori spostamento a livello di mezzo pixel.
Inoltre per ridurre la sua complessità lo standard è stato diviso in *Profiles* e *Levels* come indicato in Tabella 1. Un dato *Profile* fa riferimento alla funzionalità dello standard, la sintassi ed il tipo di algoritmi di compressione, mentre i *Levels* si riferiscono ai parametri di compressione: la risoluzione video ed il *bitrate*. Attualmente il profilo MP@ML (*Main Profile at Main Level*) è quello più sfruttato; per esempio nei DVD e nelle trasmissioni TV digitali sia via satellite che terrestri.

[11] Il formato video SIF è non interlacciato con una risoluzione di metà dello standard BT.601.

Profiles / Level	Simple (I,P)	Main (I,P,B)	4:2:2 (I,P,B)	SNR (I,P,B)	Spatial (I,P,B)	High (I,P,B)
High		4:2:0 1920x1152 1920x1080 90Mb/s				4:2:0 o 4:2:2 1920x1152 1920x1080 100Mb/s
High 1440 (HDTV)		4:2:0 1440x1152 1440x1080 60Mb/s			4:2:0 1440x1152 1440x1080 60Mb/s	4:2:0 o 4:2:2 1440x1152 1440x1080 80Mb/s
Main	4:2:0 720x576 720x480 15 Mb/s	4:2:0 720x576 720x480 15Mb/s	4:2:2 720x608 50 Mb/s	4:2:0 720x576 720x480 15Mb/s		4:2:0 or 4:2:2 720x576 720x480 20Mb/s
Low		4:2:0 352x288 352x240 4Mb/s		4:2:0 352x288 352x240 4Mb/s		

Tabella 1: Schema dei Profiles e Levels nello standard MPEG-2. Le risoluzioni in blu si riferiscono ai sistemi TV con frequenza di quadro di 30 Hz (NTSC).

La parte audio dello standard prevede un sistema di alta qualità a cinque canali: Sinistro, Destro, Centrale e due canali *Surround* o suono 3D, più un sesto canale dedicato alla riproduzione dei bassi, *subwoofer* o **LFE** (*Low Frequency Enhancement*).
Il sistema di **multiplexing** è simile a quello dell'MPEG-1, alla generazione del *Program Stream* è stato affiancato il *Transport Stream* appositamente progettato per mezzi trasmissivi suscettibili di provocare errori (*Bit Error Rate* o **BER** molto alto) come le trasmissioni satellitari. Con questo sistema si possono "mescolare" simultaneamente su uno stesso flusso di dati più programmi televisivi con i relativi canali audio.

DV

Lo schema di compressione DV è il prodotto della cooperazione di molti produttori con l'intento di fornire uno standard per le videocamere Digitali. Dopo la pubblicazione dello standard questo sistema è stato adottato velocemente con grande successo.
La compressione DV usa la trasformata DCT e la quantizzazione dei coefficienti, come in JPEG e MPEG, ma a differenza di essi è stato progettato per fornire una quantità di dati fissa per ogni frame.
Ci sono due versioni di codifica DV una a 25[12] Mbit/s con un rapporto di compressione di 5:1, per il grande pubblico, ed un'altro a 50 Mbit/s ed un rapporto di compressione di 3,3:1 per gli studi televisivi professionali.
A differenza degli standard MPEG esistono solo frame di tipo **I** (*Intraframe*), questo permette l'editing non lineare delle sequenze video, una delle caratteristiche che hanno decretato il successo della codifica DV.

[12] Per l'esatteza sono 25207200 bits/secondo

DIVX

La codifica Divx è molto importante perché ha rivoluzionato l'area della compressione video. Anche la sua storia è importante. Tutto ha origine da un **codec**[13] che la Microsoft aveva realizzato per il suo formato ASF (Advanced Streaming Format). Tale formato era nato per distribuire materiale video via WEB, nel quale era stata inserita una protezione che impediva la codifica nell'altro formato molto più diffuso: **AVI** (Audio Video Interleaved). Un gruppo di *hacker* modificò il codice di questo modulo software eliminando la protezione rendendolo compatibile con i files AVI. Il codec così modificato venne reso pubblico via Internet con il nome di **Divx 3.11**.
Con il Divx 3.11 si ottengono codifiche tanto buone da diventare un ottimo sistema per fare backup di DVD Video. Non solo, è anche diventato il principale mezzo per l'interscambio di materiale video di qualsiasi genere.
Successivamente nacquero dei gruppi di lavoro negli ambienti Open Software che hanno sviluppato e rilasciato versioni riscritte *ex-novo* (e quindi legali) di questo tipo di codec chiamati Divx 4.x, 5.x, Xvid ecc. con una buona compatibilità verso il basso. Inoltre questi codec sono in continua crescita tecnologica fino ad abbracciare i nuovi standard MPEG4 e H.264.
Normalmente i video di tipo Divx o MPEG4 sono visualizzabili solo con un computer ma la sua diffusione è stata così vasta che molte aziende del settore audio e video hanno cominciato a produrre lettori DVD in grado di leggere anche questi formati.

H.263

Nel 1995 l'ITU-T ha emesso la raccomandazione H.263 evoluzione del H.261. Tale standard è stato studiato inizialmente per bassi bitrate (e quindi per livelli elevati di compressione) tipicamente nell'intervallo da 20 a 30 Kb/s ma questa limitazione è stata superata ed il suo utilizzo comprende un'ampia gamma di formati video anche con considerevoli bitrate (Tabella 2). Come estensione dell'H.261 tiene conto, inoltre, di tutte le tecniche che hanno portato agli standard MPEG.
Le differenze più significative rispetto allo standard precedente e all'MPEG risiedono nel diverso tipo di codifica dei coefficienti dopo la trasformata DCT e nella stima più accurata dei vettori di spostamento. Inoltre è possibile negoziare con il decodificatore una serie opzioni avanzate che aumentano le prestazioni sia in termini di qualità che nell'efficacia della compressione.

Formato Video	Pixel di Luminanza (Y)	Pixel di crominanza (U,V)
Sub-QCIF	128 × 96	64 × 48
QCIF	176 × 144	88 × 72
CIF	352 × 288	176 × 144
4CIF	704 × 576	352 × 288
16CIF	1408 × 1152	704 × 576

Tabella 2: formati video permessi nella raccomandazione H.263

L'efficienza raggiunta dal complesso degli algoritmi di compressione presenti nell'H.263 ne fanno un valido strumento in varie aree applicative. Queste comprendono oltre la videoconferenza, la videotelefonia, lo streaming su Internet, la telemedicina, i sistemi di sorveglianza ed molti altri ancora
Lo standard H.263 come il suo predecessore descrive solo la parte relativa alla codifica video;

[13] Codec sta per enCODer e DECoder. Modulo Software per la codifica e decodifica video, che può essere usato da varie applicazioni

Nelle applicazioni pratiche a questi vengono affiancati l'audio, anche'esso compresso secondo altri standard (G.7xx), il quale deve essere poi integrato e sincronizzato con le immagini con l'ausilio di ulteriori standard che formano una famiglia di protocolli che coprono questi ed altri aspetti presenti in una data applicazione. Un esempio di questa famiglia di standard è l'H.320 basato su infrastrutture di comunicazione ISDN e linee telefoniche commutate e l'H.323 basato sui protocolli di rete TCP/IP (Tabella 3).

Standard	H.320	H.323
Date Approved	1990	1996
Network	Narrowband switched digital ISDN	Packet-switched networks (LAN/ WAN, ATM)
Video	H.261 H.263	H.261 H.263
Audio	G.711 G.722 G.728	G.711 G.722 G.723 G.728 G.729
Multiplexing	H.221	H.225.0
Control	H.230 H.242	H.245
Multipoint	H.231 H.243	
Data	T.120	T.120

Tabella 3: Alcuni Standard per la videoconferenza

ERRORI di Compressione

Le immagini compresse specialmente quelle con *bitrate* basso mostrano degli artefatti che ne degradano la qualità e che dipendono dagli algoritmi di compressione usati. Le distorsioni più importanti sono: il *blocking*, il *ringing*, il *blurring* e l'*effetto granulare*.

L'effetto di ***blocking*** è quello che si rileva maggiormente negli standards JPEG e MPEG. Esso è dovuto alla tecnica di partizionamento dell'immagine in blocchi, i quali sono poco correlati con quelli adiacenti. Il difetto è osservabile con l'apparizione di falsi bordi tra i vari blocchi nella quale è divisa l'immagine.
L'effetto di ***ringing*** è dovuto ad un'insufficiente quantizzazione delle alte frequenze spaziali dell'immagine. Questo provoca degli errori nelle immediate vicinanze dei bordi, osservabili come bordi spuri aggiuntivi.
Il ***blurring*** (sfocamento) è un effetto che avviene per tutti gli algoritmi di compressione di tipo lossy ed è anch'esso dovuto ad una perdita di componenti di alta frequenza dell'immagine. C'è da aggiungere che tale effetto non sempre è fastidioso all'osservatore, inoltre, con l'aumentare della distanza il difetto decresce.
La perdita delle medie frequenze spaziali nelle immagini porta ad un altro effetto di perdita di qualità chiamato ***texture deviation*** il quale si presenta come rumore granulare (effetto di granulosità dell'immagine).

Nel caso di sequenze video avremo anche degli artefatti di tipo temporale come il ***flickering*** (sfarfallio) ed il movimento a scatti o ***motion jerkiness***. Il primo è il risultato di differenze in luminosità tra frame (o parti di esso) adiacenti mentre l'altro è dovuto ad una diminuzione (a

bassi bitrate) del *frame rate* tale da scendere oltre la soglia di percezione dei movimenti fluidi (25 frame/s).

Esistono, tuttavia, varie strategie per migliorare la qualità delle immagini affette da tali difetti. Una è quella di eliminarli alla sorgente con sistemi di preprocessing (o prefiltering) quindi prima della codifica, l'altra è quella del postprocessing, dopo il decodificatore.

MPEG4 (ISO 14496)

L'MPEG4 è anch'esso uno standard sviluppato dal **M**oving **P**icture **E**xperts **G**roup organo dell'ISO ed è un evoluzione dell' MPEG1 e dell'MPEG2. Esso è diviso in numerose "parti" o sub-standard con lo scopo di regolamentare le problematiche che riguardano l'audio-video e di fornire all'industria del settore un ampia gamma di regole dalle quali attingere per i loro prodotti.
L'area che lo standard copre è molto vasto: Televisione digitale, grafica animata, Applicazioni grafiche interattive, Multimedia, WWW e relative estensioni.

I sub-standard dell'MPEG4 più interessanti sono:

ISO/IEC 14496-Part 2. Questa parte dello standard, chiamata anche "MPEG4 visual", pone l'enfasi sulla flessibilità fornendo dei mezzi per poter trattare una ampia gamma di materiale video. Oltre ai tradizionali frame video, viene introdotto il cencetto di "video object", parte di una scena di forma arbitraria, in pratica una scena video è vista come una composizione di questi oggetti. Sono inoltre previsti immagini fisse e materiale video misto con componenti naturali e artificiali (generati da computer). Queste funzioni sono fornite a mezzo di una serie "profili" i quali sono delle raccomandazioni da seguire per determinate applicazioni.

ISO/IEC 14496-Part 3. Questa parte dello standard si occupa della codifica audio la quale è stata migliorata rispetto al passato con l'introduzione dell'Advanced Audio Coding (AAC).

ISO/IEC 14496-Part 10. Questa parte chiamata anche **H.264**/**A**dvanced **V**ideo **C**oding è il più recente standard che riguarda la compressione video. Esso è stato sviluppato inizialmente dal VCEG (**V**ideo **C**oding **E**xperts **G**roup) organo dell' ITU con la sigla H.264 ed infine finalizzato con la formazione della commissione **JVT** (Joint Video Team) che comprendeva anche l'MPEG dell'ISO/IEC. Questo standard che prevede sette profili, contiene delle nuove caratteristiche che lo rendono più affidabile ed efficiente nella compressione video anche a *bitrate* bassi. La nuova generazione di media ottici HD-DVD e BLU-RAY usa questo tipo di codifica.

ISO/IEC 14496-Part 14. Questa parte dello standard insieme all'**ISO/IEC 14496-12** si occupa del formato dei *files* che hanno contenuti di tipo MPEG4.

BIBLIOGRAFIA

Watkinson, J.*, The MPEG Handbook*. Focal Press, 2001

Moammed, G., *Standard Codecs: Image Compression to Advanced Video Coding*. IEEE Telecommunications Series 49, 2003.

Symes, P. *Video Compression Demystified*. McGraw-Hill, 1998.

Richardson, Iain E.G., *H.264 and MPEG-4 Video Compression*. Wiley, 2003.

Clarke, R., *Digital Compression of still images and video*. Academic Press, 1995.

Waggoner, B., *Compression for great digital video*. CMP Books, 2002.

Solari, S.J., *Digital video and audio compression*. McGraw-Hill, 1997.

Effelsberg, W. e Steinmetz, R., *Video Compression Techniques*. Springer & Verlag, 1998.

Bhaskaran, V., *Image and Video compression standards*. 2 ed. Kluwer 1997.

Salomon, D., *Data Compression: the complete reference*. Springer 2004.