Che cosa è, come funziona: Utensili per la compressione video

ing. Marzio Barbero e
ing. Natasha Shpuza

1. Premessa

Si è visto, nella scheda precedente, che il segnale video digitale codificato secondo la Rac. ITU-R BT.601 è caratterizzato da un numero di campioni elevato: vi sono 10 368 000 elementi d'immagine (la sola porzione attiva, senza considerare i sincronismi) al secondo. Considerando il formato 4:2:2 e una codifica a 8 bit al campione per la luminanza e altrettanti per le due componenti di crominanza, il bit rate associato all'informazione video è quindi circa 166 Mbps.

Negli anni '80 si realizzarono i sistemi atti a manipolare e registrare un tale flusso di dati, ma era evidente la necessità di sviluppare tecniche per ridurre il bit-rate necessario al trasporto, poiché le capacità dei canali (in particolare i ponti radio numerici) disponibili non consentivano il trasferimento dei segnali sotto forma digitale tra i luoghi di ripresa, produzione e postproduzione.

Per ridurre il bit-rate associato al video, senza compromettere significativamente la qualità delle immagini si sviluppò e si ottimizzò un insieme di algoritmi o utensili (*tool-kit*) per comprimere il segnale video e consentirne il transito per i canali disponibili.

2. Cenni storici

I ponti radio disponibili negli anni '80 per la rete di contribuzione (collegamenti fra studi e centri di produzione) e di distribuzione primaria (per trasferire il segnale ai centri di diffusione terrestre) avevano una capacità di 34 Mbps (in Europa) e di 45 Mbps (negli Stati Uniti). Era quindi necessario definire uno standard per comprimere il segnale di almeno un fattore 6 mantenendo una qualità dell'informazione video tale da consentire eventuali operazioni di post-produzione.

Per ottenere livelli elevati sia per il fattore di compressione che per la qualità occorre utilizzare contemporaneamente tecniche (sviluppate a partire dagli anni '50) che sfruttano la ridondanza statistica dei dati, tecniche basate sulla ridondanza spaziale (la trasformata DCT era stata proposta per le immagini nel 1974 [1]) e sulla ridondanza temporale (compensazione del movimento), ottimizzando la scelta di algoritmi e parametri di codifica in funzione delle caratteristiche psicovisive umane.

Per questa opera di definizione e ottimizzazione degli algoritmi fu molto importante il contributo dei centri di ricerca e sviluppo dei radiodiffusori e dei produttori di apparati per telecomunicazioni europei, in particolare di quelli partecipanti al progetto europeo Eureka 256, fra cui il Centro Ricerche della Rai. Tale progetto mise a punto un sistema in grado di operare anche sul segnale ad alta definizione, con un fattore di compressione superiore a 10, e culminò nel 1990 con la realizzazione degli apparati che consentirono la trasmissione da parte della Rai delle partite del campionato mondiale di calcio. I risultati tecnologici ottenuti furono oggetto di numerose pubblicazioni [2-5] e riconoscimenti a livello internazionale (figura 1).

Fig. 1 - La copertina e l'indice del numero speciale di Elettronica e Telecomunicazioni del novembre 1990 dedicato alla trasmissione di alcune partite del campionato mondiale di calcio Italia '90 in alta definizione e codificate con il sistema di riduzione della ridondanza messo a punto nell'ambito del progetto Eureka 256.
Pagina del numero 2 del 1991 "Assegnazione del Montreaux Achievement Gold Medal all'ing. Marzio Barbero del Centro Ricerche della Rai (Montreux, 13-18 giugno 1991) ".

Fig. 2 - Uno dei codificatori HDTV utilizzati per Italia '90. Il codificatore è racchiuso in un sottotelaio da 19" di larghezza e 6 unità di altezza. Il consumo è pari a circa 200 W.

Per quanto riguarda la tv a definizione convenzionale, nel 1992 fu emesso lo standard europeo ETS 300 174 [6] a cui corrisponde a livello mondiale la Rac. ITU-T J.81 [7].

A partire da metà anni '90 si è poi rapidamente sviluppata la tecnologia, in particolare quella legata al sistema MPEG-2 oggetto della scheda successiva, e ciò ha comportato ulteriori miglioramenti dal punto di vista degli algoritmi, ma soprattutto dal punto di vista delle possibilità di integrazione, passando da apparati delle dimensioni di un *rack* (figura 2) a circuiti integrati caratterizzati da minimo ingombro e consumo: un dispositivo recentemente messo in commercio è costituito da un singolo *package* plastico a 273 piedini, dimensioni 15x15 mm, consumo 330 mW, è in grado di co/decodificare segnali audio/video MPEG-2.

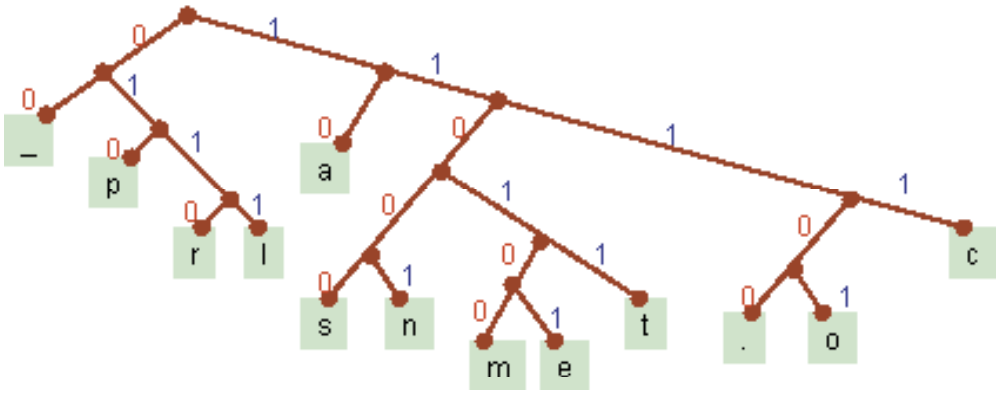

Fig. 3 - Albero binario costruito a partire dalla statistica dei simboli presenti nella stringa di esempio.

3. Ridondanza statistica e compressione *lossless*

Per ridurre il bit-rate associato ad un segnale digitale si può sfruttare la ridondanza di tipo statistico presente nel flusso di dati, utilizzando tecniche sviluppate a partire dagli anni '50 per ridurre il numero di bit associato a testi o a dati di tipo informatico.

3.1 Codice di Huffman

Uno dei metodi più conosciuti è quello proposto dal matematico D.A. Huffman nel 1952.

Questo algoritmo è applicabile quando la sorgente emette simboli caratterizzati da una probabilità non uniforme.

Per capire come funziona, utilizziamo un semplice esempio basato su un breve testo:

sopra_la_panca_la_capra_campa._sotto_ la_panca_la_capra_crepa.

E' costituito da un totale di 61 caratteri in cui si individuano 13 simboli differenti (comprendendo fra i simboli anche lo spazio e il punto).

Per codificare ciascun simbolo in binario si potrebbero utilizzare parole costituite da 4 bit (con quattro bit le configurazioni possibili sono 2^4 cioè 16) e quindi l'intera stringa sarebbe rappresentabile da 244 bit in totale.

La codifica secondo Huffman permette di ridurre il numero di bit totali associando a ciascun simbolo una parola binaria di lunghezza non fissa, tale per cui ai simboli più probabili vengano associate parole più corte, ai simboli meno probabili parole più lunghe.

L'algoritmo funziona in questo modo:

- Analizza il numero di ricorrenze di ciascun simbolo (nell'esempio *a* ricorre 16 volte, lo spazio _ 11 volte, mentre *m* ed *e* compaiono una sola volta)
- Accomuna i due elementi meno frequenti in un sottoinsieme somma (nell'esempio *m* ed *e*) e li distingue associando, ad esempio, *0* ad *m* e *1* ad *e*.

Tab. 1 - Tabella di assegnazione dei codici binari in base all'albero di figura 1.

simbolo	numero di occorrenze	probabilità di occorrenza	parola di codice
a	16	0,262	10
_	11	0,180	00
p	7	0,115	010
c	6	0,098	1111
l	4	0,066	0111
r	4	0,066	0110
o	3	0,049	11101
n	2	0,033	11001
s	2	0,033	11000
t	2	0,033	11011
.	2	0,033	11100
e	1	0,016	110101
m	1	0,016	110100

- Ripete iterativamente il processo con i due sottoinsiemi meno frequenti, usando lo stesso procedimento e considerando un tutt'unico il sottoinsieme costituito da *m* ed *e*, caratterizzato da una probabilità di occorrenza pari alla somma delle probabilità associate ai singoli *m* ed *e*.

- Si crea così un albero (figura 3) costituito da una serie di ramificazioni binarie, in cui le foglie costituite dai simboli più rari sono più lontane dalla radice e sono identificate da un codice binario più lungo.

- seguendo il percorso dalla radice fino alla foglia, si determina il codice assegnato a ciascun simbolo (tabella 1).

Il decodificatore, seguendo il percorso indicato dai bit che esamina in sequenza è in grado di individuare univocamente (un codice di Huffman non è mai prefisso di un altro) la foglia, ovvero il simbolo relativo a ciascuna parola, anche se le parole sono a lunghezza variabile (VLC, *Variable Lenght Code*).

Tornando all'esempio, la stringa di caratteri "*sopra*" diventa, codificata in binario *1100*01110*1010*0110*10.*

Nel caso di questa stringa, un codice a lunghezza fissa, assegnando 4 bit a simbolo, avrebbe richiesto 20 bit, mentre con la codifica VLC ne sono sufficienti 19. Un guadagno significativo si ha nel caso della stringa "*_la_*", che viene codificata con *00*0111*10*00 ovvero con, mediamente, 2,5 bit per carattere.

Nel complesso, la frase richiede 198 bit, anziché i 244 richiesti da una codifica a lunghezza fissa, con un risparmio dell'ordine del 19% e senza perdita di informazione (codifica *lossless*).

L'esempio è semplice e riduttivo: in genere i fattori di compressione ottenibili sono più elevati partendo da testi contenenti caratteri ASCII, codificati con parole a lunghezza fissa di 7 bit.

Il metodo funziona se codificatore e decodificatore utilizzano lo stesso albero, ovvero la stessa tabella, e ciò può essere ottenuto o adottando una tabella di assegnazione fissa, oppure inviando (nel caso di trasmissione) o memorizzando (nel caso di un file) la tabella prima dei dati compressi. E' anche possibile utilizzare l'algoritmo in modo adattativo, ovvero si parte da una tabella che viene aggiornata parallelamente dal codificatore e dal decodificatore in funzione dei simboli via via trasmessi.

3.2 Algoritmo LZW

Un ulteriore miglioramento di efficienza può essere ottenuto considerando, anziché i singoli simboli emessi dalla sorgente, insiemi di simboli. Tornando all'esempio precedente, un notevole guadagno è ottenibile utilizzando una singola parola di codice per rappresentare "*_la_*" che compare 4 volte e "*pra*" che compare 3 volte.

Questa tecnica è alla base dell'algoritmo noto come LZW (da Jacob Ziv e Abraham Lampel che pubblicarono due articoli nel 1977 e 1978 e Terry Welch che propose una modifica alle loro proposte nel 1984). Codici basati su LZW sono utilizzati, ad esempio, per la compressione di immagini (in formato GIF, *Graphics Interchange Format).* A seguito di controversie sui brevetti alla base della tecnica LZW e quindi GIF, è stato successivamente definito il formato PNG (*Portable Network Graphics*).

3.3 Run-Length Encoding (RLE)

Questo tipo di codifica si basa sulla ripetizione all'interno del messaggio di uno stesso simbolo. In questo caso è possibile, ad esempio codificare l'occorrenza di n simboli uguali con due byte, il primo, denominato *run count*, indica il numero di ripetizioni ed il secondo, denominato *run value*, indica il valore (ad esempio il codice ascii di un carattere, o il livello di luminanza di un pixel).

Fig. 4 - Confronto fra le dimensioni relative di un'immagine SDTV (*Standard Definition TV*) da 720x576 elementi di immagine, quella CIF (*Common Intermediate Format*, 352x288), QCIF (*Quarter CIF., 176x144*), QQCIF (88x72).

4. Irrilevanza e compressione *lossy*

Le tecniche precedentemente descritte sono efficaci nel caso di compressione di immagini grafiche, ed in effetti sono alla base di tutti i formati utilizzati per le immagini per PC, soprattutto quando vi sono sequenze di pixel uguali fra loro o strutture ripetitive, ma non consentono di ottenere elevati fattori di compressione nel caso di immagini di tipo naturale.

Per ottenere fattori di compressione superiori è necessario accettare la perdita di informazione (sistemi *lossy*), ovviamente riducendo al minimo la percezione dei difetti introdotti.

4.1 Ridondanza spaziale

Per ridurre il bit-rate finale un primo approccio è quello di ridurre il numero di campioni dell'immagine da codificare. In figura 4 sono messe a confronto le dimensioni relative di un'immagine da 720 elementi di immagine per riga e 576 righe, come previsto dalla Rac. ITU-R BT.601, con il formato CIF, utilizzata in applicazioni di videoconferenza e in MPEG-1. Il formato QCIF è spesso utilizzato per il webcasting, mentre il QQCIF è impiegato per videotelefonia (UMTS). La riduzione della definizione spaziale, orizzontale e verticale, è senz'altro efficiente e può essere efficace per immagini relative al voltro umano, o a singoli oggetti o animali (un fiore o una farfalla), ma non è utilizzabile in campo radiodiffusivo, a causa della difficoltà nel riprodurre i dettagli.

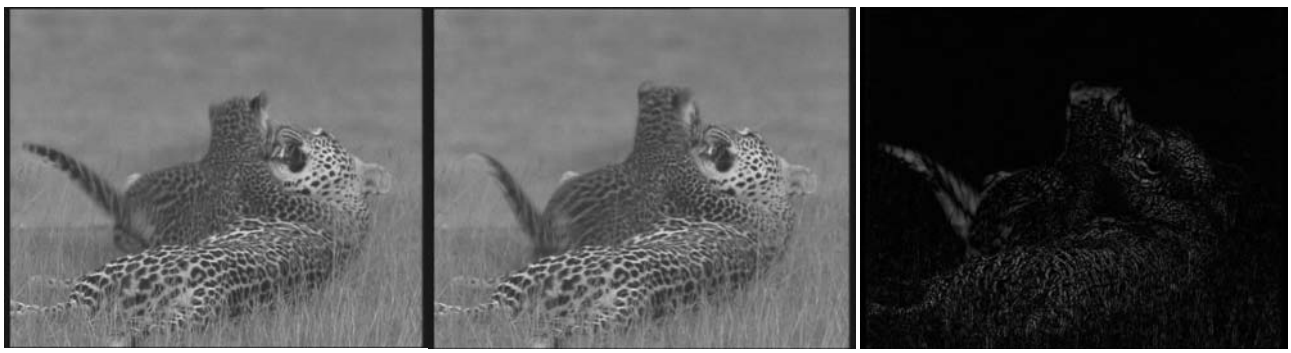

Fig. 5 - Due quadri successivi (sola luminanza) di una sequenza video e immagine differenza fra le due.

4.2 Ridondanza temporale

Le sequenze video sono caratterizzate da un'altra forma di ridondanza, quella temporale. Nel sistema europeo vi sono 25 quadri (ciascuno composto da due semiquadri interlacciati) al secondo. La risoluzione temporale è generalmente ridotta nelle applicazioni menzionate precedentemente: nel formato CIF di fatto si elimina un semiquadro su due, ma spesso il numero di quadri è ulteriormente ridotto (ad esempio 12 o 6 quadri al secondo). Ovviamente anche in questo caso l'informazione eliminata non è irrilevante ed il degradamento è sensibile.

Un ulteriore modo per sfruttare la ridondanza temporale è quello di codificare le differenze fra quadri successivi, trasmettendo o memorizzando solo l'informazione che cambia, da un quadro a quello successivo. Dalla figura 5 appare evidente che spesso le differenze fra due quadri successivi sono piccole, molto prossime al valore 0: è quindi possibile sfruttare tecniche di riduzione della ridondanza statistica, sfruttando la distribuzione non uniforme delle ampiezze delle differenze (i campioni video di luminanza sono invece distribuiti uniformemente fra il nero e il bianco).

4.3 Compensazione del movimento

I campioni co-posizionati del quadro precedente possono quindi costituire una predizione molto buona per i campioni da codificare. Nel caso che gli oggetti ripresi si muovano all'interno della scena oppure la telecamera effettui movimenti orizzontali (*panning*) o verticali (*tilting*) è possibile ottenere una buona predizione se si riesce a stimare il movimento e a codificare sia la predizione che il vettore movimento che permette al decodificatore di individuare l'informazione già trasmessa (parte del quadro precedente) da cui si è ottenuta la predizione. Un'ottima predizione è possibile associando un vettore movimento a ciascun campione d'immagine, ma ciò richiede la trasmissione di un enorme numero di vettori; si potrebbe trasmettere un solo vettore movimento, globale per l'intera immagine, ma in tal caso sarebbe trascurabile il guadagno di predizione: il compromesso in genere consiste nell'organizzare l'immagine in blocchi (ad esempio di 8x8 o 16x16 campioni) ed individuare e trasmettere un solo vettore movimento per ciascun blocco (figura 6).

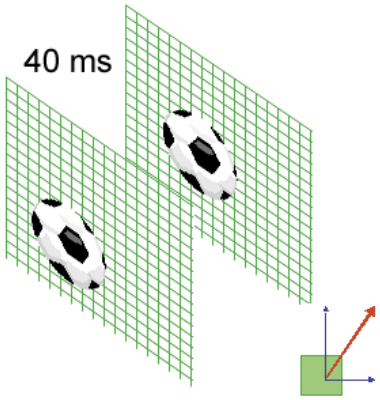

Fig.6 - Un oggetto si trova in due posizioni differenti in due quadri successivi: un vettore, caratterizzato dalle due componenti x e y può consentire di compensare la traslazione dei pixel corrispondenti ad un oggetto che nei 40 ms intercorsi fra i due quadri appare spostato spazialmente, nell'immagine.

4.4 Quantizzazione

Una riduzione del bit-rate è ottenibile quantizzando con una minor precisione i campioni (figura 7): la perdita di informazione (indicata come errore di quantizzazione o rumore di quantizzazione) è rilevante e proporzionale al numero di bit al campione risparmiati.

Fig.7 - la riduzione del numero di bit per campione dà origine all'incremento del rumore di quantizzazione: in questa immagine di sola luminanza si ha un effetto di solarizzazione crescente a partire da quella in alto a sinistra (8 bit) a quella in alto a destra (6 bit), in basso a sinistra (4 bit), in basso a destre (2 bit).

5. Una trasformazione di dominio

Di fatto tutti i metodi precedentemente descritti sono utilizzati nei sistemi di compressione, sia quello sviluppato in Eureka 256, sia in quelli MPEG.

Non vengono però applicati direttamente ai campioni video perché, come si è visto, ad elevati fattori di compressione corrisponderebbero altrettanti elevati fattori di distorsione e perdita in qualità dell'immagine. Prima di essere compressi i campioni video vengono raggruppati (in genere in blocchi 8x8) e trasformati.

La trasformazione normalmente adottata è la DCT (*Discrete Cosine Transform*). La DCT è un algoritmo matematico che può essere descritto in molti modi: moltiplicazione matriciale, rotazione d'asse in uno spazio a 64 dimensioni, FFT (*Fast Fourier Transform*) di un blocco accanto al suo blocco riflesso, blocco decomposto nelle sue funzioni basi.

Comunque la si consideri, lo scopo della Trasformata Coseno Discreta è quello di ottenere, a partire dal blocco di 8x8 campioni video (8 bit per campione), un altro blocco di 8x8 valori, (generalmente rappresentati da 11 o 12 bit): a questo punto del processo il numero di bit associato a ciascun blocco è aumentato, ma la distribuzione statistica dei valori è radicalmente modificata.

Le funzioni basi rappresentano le frequenze spaziali, dalla continua alle frequenze orizzontali, verticali e diagonali più elevate (figura 8).

Il sistema psicovisivo umano considera meno rilevanti le frequenze spaziali più elevate e quindi si può applicare di fatto una riduzione di banda: i coefficienti sono pesati dividendoli per costanti di peso diverso a seconda della posizione nella matrice (figura 9).

Nel caso ciò sia conveniente, i blocchi trasformati non sono quelli contenenti i campioni video, ma quelli ottenuti come differenza, a partire da valori di predizione determinati anche sfruttando la compensazione del movimento.

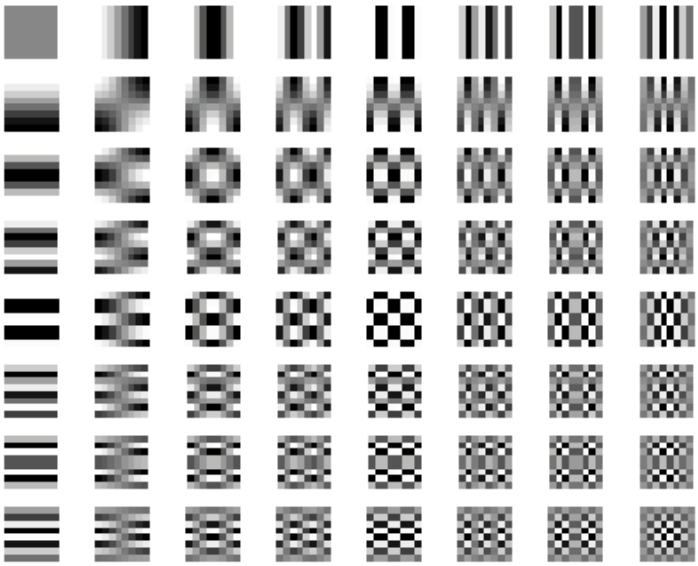

Fig.8 - Le 64 funzioni base nel caso della trasformata DCT 8x8. Qualsiasi blocco di campioni video può essere rappresentato da una combinazione di alcune delle funzioni opportunamente pesate. In genere è presente la prima funzione in alto a sinistra, che rappresenta la componente continua (il valore medio del blocco) più un certo numero di funzioni, con una predominanza di quelle corrispondenti alle frequenze spaziali più basse, cioè disposte nell'angolo in alto a sinistra.

Fig.9 - La matrice di pesatura è utilizzata per dividere i coefficienti per un valore legato al rumore di quantizzazione che il sistema psicovisivo è in grado di tollerare in funzione delle singole frequenze spaziali: il minimo di rumore è tollerato per la componente continua (cioè per i blocchi e, in definitiva, le immagini in cui ci sono lente variazioni di luminanza), mentre un maggiore errore di quantizzazione è accettabile per blocchi rumorosi o ricchi di dettagli. Questa è la matrice di pesatura adottata per MPEG-2.

8	16	19	22	26	27	29	34
16	16	22	24	27	29	34	37
19	22	26	27	29	34	34	38
22	22	26	27	29	34	37	40
22	26	27	29	32	35	40	48
26	27	29	32	35	40	48	58
26	27	29	34	38	46	56	69
27	29	35	38	46	56	69	83

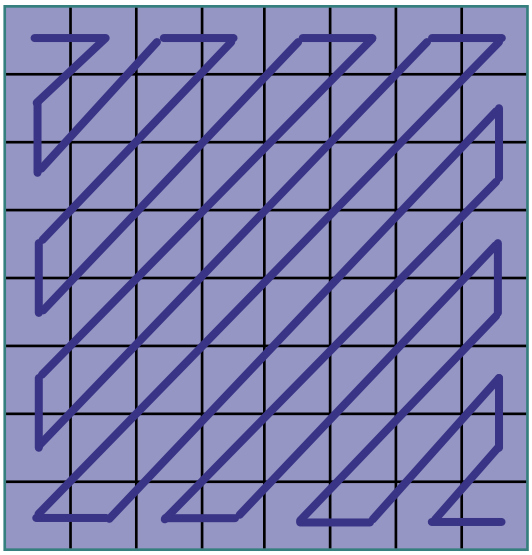

Fig.10 - I coefficienti moltiplicativi delle frequenze base sono trasmessi seguendo un ordine secondo un percorso a zig-zag: dapprima il coefficiente relativo alla componente continua e via via quelli relativi alle frequenze spaziali più elevate.

I coefficienti vengono riordinati, prima della trasmissione, secondo un percorso a zig-zag (figura 10) I coefficienti sono codificati mediante VLC, sfruttando così la statistica non uniforme delle loro ampiezze, mentre le lunghe sequenze di zeri che si vengono a creare possono essere codificate con RLE. Poiché in genere gli ultimi coefficienti sono nulli, essi non vengono trasmessi, ma l'ultimo *run-length* è, qualunque sia la sua lunghezza, sostituito da un unico simbolo, denominato EOB (*End of block*).

La lunghezza totale del blocco codificato è quindi variabile e dipende dalla configurazio-

Grazie alla ridondanza temporale, i coefficienti DCT, soprattutto quelli relativi alla componente continua e alle basse frequenze spaziali, hanno ampiezza inferiore a quelli che si otterrebbero trasformando direttamente i campioni video.

I 64 coefficienti moltiplicativi delle funzioni base assumono quindi valori prossimi allo zero e, anche grazie all'effetto della matrice di pesatura, sono spesso nulli quelli posizionati verso l'angolo inferiore a destra (frequenze spaziali più elevate).

Fig.11 - Se si suddivide un'immagine (la trattazione si riferisce alla luminanza, ma può essere estesa alle componenti di crominanza) in blocchi 8x8, questi hanno in genere caratteristiche diverse. Ad esempio il blocco in alto è tratto da una porzione del viso e si notano minori variazioni (i campioni sono più correlati), il blocco intermedio è tratto dalla siepe ed è caratterizzato da variazioni più ampie della luminanza (è una struttura casuale, simile al rumore), il blocco in basso è relativo alla collana e quindi rappresentativo di contorni e dettagli ben definiti.
Ciascun blocco, applicando la DCT, è rappresentabile come combinazione delle funzioni base.
Il numero di coefficienti generato è minore per il blocco superiore, il rumore di quantizzazione è più percepibile nel caso blocco inferiore rispetto a quello intermedio.

ne del blocco dei campioni video che lo ha originato.

I blocchi 8x8 presentano livelli di difficoltà differenti per la compressione (figura 11). I blocchi che danno origine ad un numero inferiore di bit sono quelli in cui vi è una forte correlazione, sono anche quelli su cui è più visibile l'eventuale rumore di quantizzazione. I blocchi che hanno una struttura più complessa, simile al rumore, sono quelli che danno origine ad un numero di coefficienti e, in ultima analisi, di bit superiore, ma sono anche quelli in cui è accettabile un maggior rumore di quantizzazione (grazie alle caratteristiche del nostro sistema psicovisivo) e quindi è possibile quantizzare più grossolanamente i coefficienti, riducendo di conseguenza il numero di bit da trasmettere. Infine i blocchi corrispondenti a contorni sono quelli più critici, danno origine a molti coefficienti non nulli ed il rumore di quantizzazione è percepibile.

Il decodificatore opera la trasformazione inversa, ricostruendo i campioni video (o le differenze) relative a ciascun blocco.

Il sistema è complesso, ma sfrutta in modo efficiente le ridondanze temporali, spaziali e statistiche del segnale video, minimizzando il degradamento percepito.

Bibliografia

1. N. Ahmed, T. Natarajan, K.R.Rao: "Discrete Cosine Transform" IEEE Trans. Computers, Vol. C-23, Jan. 1974, pp. 90-93

2. Speciale Italia '90, Elettronica e Telecomunicazioni, No. 3, 1990.

3. M. Barbero, S. Cucchi, M. Stroppiana: "A Bit-Rate Reduction System for HDTV Transmission", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 1, No. 1, Marzo 1991, p. 4.

4. M. Barbero, M. Hofmann, N. D. Wells: "DCT Source Coding and Current Implementation of HDTV", EBU Tech. Review, No. 251, Spring 1992, p. 45-54

5. M. Barbero, M. Stroppiana: "Video Compression Techniques and Multilevel Approaches", SMPTE Journal, Vol. 103, No. 5, maggio 1994, p. 335.

6. ETS 300 174: "Network Aspects (NA); Digital coding of component television signals for contribution quality applications in the range 34 - 45 Mbit/s" (Nov. 1992).

7. Recommendation ITU-T J.81: "Transmission of component-coded digital television signals for contribution-quality applications at the third hierarchical level of ITU-T Recommendation G.702" (1993).