“Università del Piemonte Orientale Amedeo Avogadro”

Corso di Laurea Specialistica in
Informatica dei sistemi avanzati e dei servizi di rete

Corso di Analisi Dati Intelligente

Studio introduttivo alle tecniche di Video Retrieval

Vicino Guido

Anno Accademico
2005/2006

Indice

1 Introduzione

L'obiettivo dei sistemi content-based retrieval (CBR), nel contesto delle immagini e dei video, è quello di recuperare un insieme di immagini o videoclip da una larga collezione selezionati sulla base di contenuti interni desiderati, in aggiunta anche ad associazioni alfa-numeriche, parole chiavi ed attributi.

L'interesse verso questi sistemi è sempre stato crescente da parte di gruppi ed aziende che si occupano di pubblicità, intrattenimento, previsioni metereologiche, telecomunicazioni, robotica, medicina, e sicurezza nazionale.

Quello che differenzia il trattamento dei video rispetto a quello delle immagini, sono principalmente due fattori, la componente sonora, ma soprattutto l'introduzione della dimensione temporale. Il flusso temporale implica una nuova dimensione di studio del contenuto, e lo scorrere dei vari fotogrammi diventa rilevante per la comprensione semantica dell'intero documento. L'analisi di un filmato si compone quindi di due compiti: la divisione temporale dello stesso in unità singole, e l'estrazione dei contenuti da queste unità, andando ad osservare contemporaneamente audio e video.

Attualmente i sistemi commerciali conosciuti che si occupano di recupero di filmati, non sono di questa tipologia ma si basano su proprietà “testuali” associate al documento digitale, quali il nome del file, alcune proprietà e caratteristiche interne del formato video, ma anche il contesto in cui vengono inserite(ad esempio il testo di una pagina Internet in cui il documento è inserito). Alcuni di questi sistemi sono comparsi come applicazioni web, orientate all'intrattenimento, i due esempi di maggior successo sono GoogleVideo e YouTube.

Una ricerca semantica non può essere condotta utilizzando le interfacce comunemente usate, ma si richiederà di fornire un sistema efficiente adatto alla formulazione di interrogazioni, come la seguente:

“Una sequenza video di sessanta secondi contenente un ragazzo in motocicletta rossa percorre una strada di città, in ambiente notturno”

E' facile osservare che un interrogazione simile non è ovviamente di facile formulazione tramite descrizioni testuali in quanto si presta a diverse interpretazioni soggettive.

Altre volte si richiederà al sistema la possibilità di essere consultato fornendo un filmato d'esempio simile a quello che vogliamo trovare, o ancora permettendo all'utente di creare lui stesso un modello. Questo modello potrà essere un disegno o un abbozzo dell'immagine da recuperare, oppure una serie di immagini ridotte all'essenziale per il recupero di un filmato.

Nel nostro studio, abbiamo deciso di fornire una breve introduzione delle principali problematiche e tecniche legate al Content Based Video Retrieval, andando dove possibile a cercare applicazioni pratiche realizzate in questo campo di ricerca.

Le tecniche usate in questo campo di studi sono diverse ed in continua evoluzione, verrà fatta una panoramica sulle più rilevanti e conosciute. L'estrazione di un documento video consta principalmente di due passi. Il primo passo è segmentazione temporale, ossia l'identificazione dei segmenti significativi componenti il documento quali shot, episodi e video. Il secondo passo è l'analisi dei contenuti vera e propria, cioè l'identificazione degli attributi che caratterizzano le regioni, gli oggetti ed il movimento di un filmato. Descriveremo inoltre altri aspetti rilevanti quali la scelta dell'interfaccia da usare, la formulazione dell'interrogazione e la creazione di indici. Incominciamo con un breve ripasso delle tecniche utilizzate per distinguere tra immagini differenti o meno; queste tecniche vengono ereditate da diversi campi di studio ma si rilevano importantissime per differenziare i vari fotogrammi di un qualsiasi documento video.

2 Analisi di similarità dei fotogrammi

Nei sistemi di video retrieval si prendono in prestito diverse tecniche prese dal retrieval su immagini. Dati due fotogrammi evidenziarne le possibili differenze o similarità è essenziale, questo ci permette ad esempio dato un filmato d'esempio di andare a cercarne altri dallo stesso contenuto all'interno di un'ipotetica base di dati, oppure ci permette di distinguere scene simili tra loro all'interno delle stesso video. La similarità quindi è rilevante per un gran numero di problemi inerenti allo studio automatico dei filmati. Esistono diverse tipologie di similarità tra i fotogrammi e vanno a studiare i seguenti fattori:

- Colore
- Trama o tessitura
- Forma

2.1 Similarità per colore

Il sistema visivo umano percepisce il *colore* attraverso differenti aspetti fisici, neuro-fisiologici e psicologici. Più precisamente la percezione del colore dal punto di vista fisico è data dalla quantità e tipologia dell'energia spettrale che colpisce la retina. L'occhio cattura l'informazione luminosa presente nel campo visivo e demanda al cervello una sua interpretazione a seconda dell'esperienza che esso ha accumulato.

Questa energia luminosa viene espressa analiticamente tramite una funzione della lunghezza d'onda

E(λ), dall'analisi di questa è possibile riconoscere le varie tonalità cromatiche appartenenti alla banda visibile dall'uomo.

Tramite tre diverse tipologie di cellule della retina l'occhio umano può percepire tre lunghezze d'onda differenti corrispondenti ad altrettanti tre colori:

- Lunghezze d'onda lunghe, corrispondenti al colore rosso.
- Lunghezze d'onda medie, corrispondenti al colore verde.
- Lunghezze d'onda corte, corrispondenti al colore blu.

L'energia spettrale della luce visibile denominata *C(λ)* viene percepita da queste cellule (dette coni) producendo segnali elettrici α_i mediante i quali viene prodotta nel cervello la sensazione del colore. Questi segnali sono descritti tramite la formula seguente:

$$\alpha_i(C) = \int_{\lambda_{min}}^{\lambda_{max}} S_i(\lambda) C(\lambda) d\lambda$$

Formula 1

Un altra componente importante importante è quello psicologica, noi relazioniamo la percezione del colore a diversi aspetti, misurati quantitativamente, questi descrivono le caratteristiche del colore percepito quali brillanza, chiarezza tinta e saturazione. Queste caratteristiche vengono descritte tramite modelli geometrici approssimati alla sensazione cromatica del nostro cervello. Nei sistemi CBIR e CBVR questi modelli matematici vengono utilizzati per la rappresentazione del colore e per il suo studio. I diversi colori vengono trattati come vettori di uno spazio n-dimensionale (tipicamente con *n=3*), due esempi di questi spazi sono quello *RGB (Red, Green, Blue)* e quello *CMY (Cyan, Magenta, Yellow)*. In questi spazi le differenze tra colori vengono valutate misurando la distanza tra i loro punti rappresentativi. L'analisi viene condotta sulla distribuzione di colore dell'immagine stabilendo una rappresentazione e misurazione opportuna, normalmente questa rappresentazione è data dall'istogramma di colore dell'immagine. L'analisi del colore permette una parziale affidabilità anche nel caso di cambiamenti di luce, cambiamenti d'angolo e scala. Un metodo comune è quello di fare confronti tra le regioni da esaminare, andando a confrontare le due distribuzioni di colore, ad esempio nello spazio *LUV (Luminanza e crominanza):*

$$C_d = \sqrt{(L_q - L_t)^2 + 4(U_q - U_t)^2 + 4(V_q - V_t)^2}$$

Formula 2

dove C_d è la distanza Euclidea pesata dei colori nello spazio mentre i pedici *q* e *t* si riferiscono alle due regioni da esaminare. Le analisi di similarità nel contesto del recupero di immagini introducono

diversi problemi in quanto ignorano totalmente l'aspetto spaziale, e quindi andranno integrate con altri fattori.

2.2 Similarità per trama

Il termine *trama* deriva dall'inglese *texture* (tessitura) e viene utilizzato per indicare delle zone dell'immagine caratterizzate da specifici "disegni", come ad esempio le foglie di un albero, la sabbia di una spiaggia o la tappezzeria di una stanza. Il calcolo della similarità tra *texture* permette una migliore discriminazione tra oggetti o regioni dell'immagine cromaticamente simili, ma semanticamente differenti, come ad esempio distinguere tra cielo e mare, o tra erba e foglie. Questa analisi viene condotta tramite decomposizione effettuando studi sulla luminanza della texture in maniera da evidenziare le variazioni di luminosità tra i vari pixel costituenti la trama. Questi studi di distanza ci permettono di costruire dei descrittori relativi a modelli psicologici quali ad esempio granulosità, direzionalità, ripetitività, periodicità e contrasto. Oppure ci permettono di trattare modelli statistici e matematici quali la trasformata *wavelet*, quella di Fourier o tramite le matrici di co-occorrenza. Nel sistema di retrieval *VideoQ* si utilizza ad esempio come funzione di distanza, la distanza Euclidea pesata lungo ciascuna caratteristica della trama relativamente alle loro varianze:

$$T_d = \sqrt{\frac{(\alpha_q - \alpha_t)^2}{\sigma_\alpha^2} + \frac{(\beta_q - \beta_t)^2}{\sigma_\beta^2} + \frac{(\theta_q - \theta_t)^2}{\sigma_\theta^2}}$$

Formula 3

dove α, β e θ si riferiscono alla ruvidezza, al contrasto e all'orientamento mentre le varie $\sigma_{\alpha,\beta,\theta}$ si riferiscono alle varianze delle caratteristiche corrispondenti. Diversi esperimenti hanno mostrato che queste componenti indipendenti si allineano molto bene alla percezione naturale delle similarità tra tessiture.

2.3 Similarità per forma

La *forma* di un oggetto viene rappresentata tramite un insieme di caratteristiche globali differenti, quali l'area, l'allungamento, la compattezza, la direzione degli assi, oppure tramite caratteristiche locali quali angoli, punti e segmenti caratteristici. Anche in questo caso nei sistemi informatici si utilizzano vettori n-dimensionali relativi allo spazio delle caratteristiche considerate. In tal maniera è possibile nuovamente effettuare studi di similarità tramite analisi di distanza tra queste caratteristiche. Altri approcci invece vanno a calcolare dati due oggetti presi in esame A e B il costo in complessità per passare da uno all'altro tramite una trasformazione $T: A \rightarrow B$. Questo tipo di studio è alla base dei sistemi come *VideoQ* che fanno uso di interrogazioni tramite *sketch*, che

andremo a descrivere in seguito. Infine un ultimo metodo consiste nell'analisi delle forme tramite semplice proprietà geometriche, quali area, proiezioni ortogonali e compattezza.

3 Segmentazione temporale

Dopo aver brevemente descritto i tre aspetti che ci permettono di andare ad analizzare le singole immagini, andiamo a descrivere il processo di segmentazione temporale di un filmato. L'intento è quello di suddividere il flusso audio e video in sequenze elementari più semplici fino ad arrivare al singolo fotogramma, cioè alla singola immagine. Questa suddivisione ci permette di semplificare il problema e di poter applicare successivamente tecniche per l'estrazione dei contenuti.

Si hanno tipicamente tre livelli di analisi: *key, shot* e *scene.*

- Un *frame* è un singolo fotogramma estratto dalla sequenza finita che costituisce l'intero video.
- Lo *shot* è una sequenza di fotogrammi ripresi in maniera continuativa dalla videocamera.
- Una *scena* è una collezione di uno o più shot riguardanti uno o più oggetti di interesse.

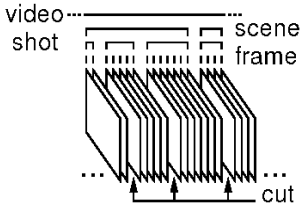

Figura 1: Struttura di un video

Alcune forme di rilevamento delle transizioni tra i vari shot sono state incorporate nell'hardware di alcune periferiche per il cinema fin dai primi anni 70, ma i ricercatori hanno iniziato a definire bene i problemi su cui lavorare solo dall'inizi degli anni 90.

La difficoltà del problema consiste nel separare i vari fattori relativi alla transizione delle immagini, che sono:

- *Movimento:* in esso sono inclusi i movimenti degli oggetti e quelli apparenti dovuti al lavoro indotto dalla telecamera.
- *Luminosità:* i cambiamenti di luce ed i rumori su di essa.
- *Shot change:* i quali possono essere bruschi o progressivi

3.1 Segmentazione di flussi non compressi

Esistono due tipologie di metodi proposti per trattare questi problemi:

- Metodi che vanno ad analizzare le differenze statistiche tra le varie cause di transizione dell'immagine.
- Metodi di segmentazione basati su modelli.

La transizione tra uno shot ed un altro può avvenire in maniera diversa. La più semplice transizione tra due shot viene definita *taglio (cut)* e consiste in un cambiamento brusco degli shot occorrente tra due fotogrammi consecutivi. Per rilevare questi cambiamenti normalmente si fa uso di confronti tramite un istogramma dei colori. La seguente Formula mostra la computazione della distribuzione tra due fotogrammi:

$$\sum_j |H_{t+1}(j) - H_t(j)|$$

Formula 4

dove *H(j)* rappresenta il *j-esimo* intervallo (bin) all'interno dell'istogramma di luminosità. Viene anche utilizzato il test del chi-quadro per confrontare due distribuzioni. La formula 5 mostra l'applicazione del test per studiare due fotogrammi consecutivi ai tempi *t* e *t+1*:

$$\sum_j \frac{(H_{t+1}(j) - H_t(j))^2}{H_{t+1}(j) + H_t(j)}$$

Formula 5

Un'altra proposta è stata quella di utilizzare i rapporti di verosimiglianza (Likelihood). Questa tecnica prende due fotogrammi e suddivide ciascuno di essi in blocchi e calcola il rapporto di verosimiglianza tra i blocchi nel primo frame ed i corrispondenti blocchi nel frame successivo:

$$L_i = \frac{\left[\frac{\sigma_t^2 + \sigma_{t+1}^2}{2} + \left(\frac{\mu_t - \mu_{t+1}}{2}\right)^2\right]^2}{\mu_t^2 \mu_{t+1}^2}$$

Formula 6

dove μ ed σ sono rispettivamente il valore medio e la deviazione standard dell'istogramma dei colori, questa tecnica effettua quella che viene chiamata una *block-based comparison*.

Il confronto tra luminosità è colore sfrutta proprietà globali del video, un altro livello di analisi è possibile andando a guardare le proprietà locali di un filmato, cioè l'analisi del cambiamento dei bordi degli oggetti. Durante una transizione netta il contenuto di un fotogramma cambia

sostanzialmente nella disposizione dello stesso, questo si traduce in un cambiamento dei contorni degli oggetti e nella loro posizione. L'idea è quella di osservare il numero dei "pixel uscenti" e di quelli "entranti" tra i due frame, ossia nel calcolare la distanza tra i pixel uscenti presenti nel fotogramma f_t ma non presenti in f_{t+1}, ed i pixel entranti presenti in f_{t+1} ma non presenti in f_t. Più formalmente questo si ottiene calcolando la seguente distanza:

$$D(i,i+1)=\frac{\sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{c}\left|P_i(x,y,c)-P_{i+1}(x,y,c)\right|}{X*Y}$$

Formula 7: Distanza tra due immagini a colori

dove i e $i+1$ sono due fotogrammi successivi di dimensione X per Y, $P_i(x,y)$ è il valore d'intensità del pixel alle coordinate x, y nel frame i, mentre c è l'indice delle componenti RGB, analogamente per $P_{i+1}(x,y)$ sul frame successivo. Se D supera una soglia predefinita si ipotizza una transizione, questa tecnica molto semplice prende il nome di *pixel comparison.*

La rilevazione del taglio è relativamente facile in quanto essa si presenta come una transizione netta, esistono però tipologie più complesse dove la transizione tra shot è *graduale*, si dividono in tre categorie: il *fade*, il *dissolve* e la *wipe.*

Nel caso del *fade* vi è un cambiamento lento nella luminosità dell'immagine che di solito inizia o risulta in un fotogramma di colore unitario come il nero (Figura 2). Si ha invece una *dissolvenza* quando le immagini del primo shot variano di luminosità in maniera tale che quelle del secondo shot si sovrappongono sopra a quelle del primo (Figura 3). Infine si ha la *pulitura* quando le immagini del secondo shot rimpiazzano quelle del primo shot attraverso un movimento lineare e continuativo.

La segmentazione nel caso di *transizioni graduali* avviene per lo più attraverso analisi cumulative nel tempo, oppure tramite analisi combinate su più caratteristiche.

Figura 2: Fade out seguito da fade in

Figura 3: Dissolvenza

Nel primo caso si effettua un analisi di variazione su intervalli temporali differenti, individuando possibili transazioni che hanno effetto su un arco variabile di tempo. Altri indicatori possono essere variazioni di luminosità piccole ma continue che possono essere indicatrici di transizioni graduali. Per rilevare tali variazioni si effettua un confronto con due soglie di variazione (*twin tresholding mechanism*) prendendo in esame le differenze cumulative tra i fotogrammi della transizione graduale. Al primo passo è utilizzata una soglia più alta T_h per rilevare i tagli, come mostrato in *Figura 4a*. Al secondo passo è impiegata una soglia più bassa T_l per rilevare il possibile fotogramma di partenza F_s di una transizione graduale e questo viene confrontato con i frame seguenti come in *Figura 4b*.

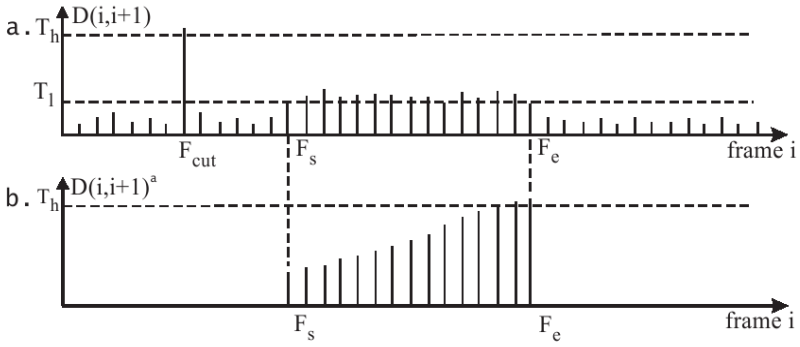

Figura 4: Twin comparison

Questa differenza viene chiamata *accumulated comparison* poiché questo valore aumenta durante una transizione graduale. Il fotogramma finale F_e della transizione viene rilevato quando la differenza tra due fotogrammi consecutivi decresce meno che T_l, mentre l'accumulated comparison è cresciuto verso un valore più alto di T_h. Se la differenza consecutiva cade sotto T_l prima che la differenza accumulata superi T_h allora il possibile fotogramma iniziale F_s è scartato e la ricerca continua per altre transizioni graduali. Alcune transazioni graduali possono sfuggire a questo controllo, in quanto la differenza cumulativa viene a trovarsi sotto la soglia più bassa. Questo

problema può essere facilmente risolto specificando un valore di tolleranza che permetta un certo numero di fotogrammi consecutivi con valori di differenza bassi prima di rifiutare una transizione candidata. Questo metodo quindi permette di rivelare sia i cambiamenti bruschi che quelli progressivi e da studi sperimentali sembra comportarsi in maniera molto buona.

Le tecniche presentante fino ad ora partono dai dati per arrivare ad effettuare la segmentazione con una metodologia *bottom-up*, esistono però anche metodi che invece usano l'approccio inverso cioè *top-down* e sono basati su modelli matematici dei dati video. Alcuni di questi metodi presentano ad esempio un identificazione basata sui modelli matematici dei processi di produzione video e di montaggio, fornendo una classificazione per le varie modifiche del video(cut, fade, dissolve). Ad esempio le transizioni graduali come *fade* e *dissolve* possono essere rilevate da modifiche cromatiche attraverso il seguente modello:

$$S(x,y,z)=S_1(x,y,t)(1-t/l_1)+S_2(x,y,t)(1-t/l_2)$$

Formula 8

dove $S_1(x,y,t)$ e $S_2(x,y,t)$ sono i due shot che subiscono la modifica, mentre $S(x,y,z)$ sarà lo shot risultante, infine l_1 ed l_2 indicano il numero di fotogrammi coinvolti nei due shot. Si cerca quindi di eseguire una tassonomia sfruttando le varie caratteristiche evidenziate dal modello.

Abbiamo presentato una breve panoramica dei metodi proposti in letteratura, altri sono stati presentati, che utilizzano teorie probabilistiche o modelli stocastici, presentiamo brevemente per concludere questa parte sulla segmentazione di video non compressi un metodo che sfrutta le *Catene di Markov Nascoste (HMM)*.

Prima di descrivere questo modello ricordiamo che l'*HMM* è un modello statistico dove si ha un processo Markoviano con parametri sconosciuti, e si vuole quindi determinare questi parametri tramite la loro osservazione.

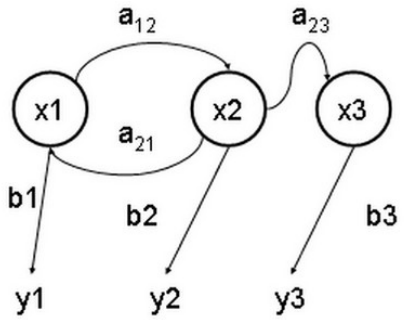

Figura 5: Esempio di diagramma di una HMM

In questo modello *stati diversi* sono usati per modellare shot, cut, fade, dissolve e altre transizioni. Gli *archi* tra gli stati modellano i possibili passaggi tra essi, ad esempio da uno *stato shot* è possibile andare in un qualsiasi *stato transizione*, ma da uno stato transizione è soltanto possibile tornare ad uno stato shot. Gli archi che partono da uno stato ed arrivano a se stesso modellano invece la lunghezza temporale del video (Figura 5).

Esistono tre tipologie di *caratteristiche* di studio usate e sono *immagine*, *audio* e *movimento* queste sono matematicamente corrispondenti a:

1. Una distanza tra fotogrammi adiacenti tramite un istogramma in bianco e nero.
2. Una distanza audio basata sulla differenza acustica negli intervalli prima e dopo i fotogrammi
3. Una stima del movimento degli oggetti tra i due fotogrammi.

I parametri dell'HMM hanno associate le probabilità di transizione per gli archi e le distribuzioni di probabilità delle caratteristiche per gli stati. Questi parametri, inizialmente sconosciuti, sono appresi e memorizzati in vettori di caratteristiche suddivisi per le varie classi di transizione tramite l'algoritmo di Baum Welch. Una volta che questi parametri sono appresi, la segmentazione del video avviene tramite l'algoritmo di Viterbi, una tecnica standard di riconoscimento nelle catene markoviane nascoste. L'algoritmo è stato testato su diverse basi di dati video e ha mostrato un miglioramento nell'accuratezza della segmentazione video rispetto ai classici approcci basati su soglia.

3.2 Segmentazione video e compressione digitale MPEG

Mentre nei precedenti approcci abbiamo trattato la segmentazione di documenti video non compressi, ora andremo a fare una breve panoramica degli algoritmi che operano sui documenti codificati tramite la codifica MPEG.

E' desiderabile poter lavorare direttamente sul flusso di dati compresso perché ormai la maggior parte dei media utilizza queste tecniche, ed inoltre perché portano a due vantaggi.

Il primo vantaggio è quello di non dover decodificare e re-codificare il documento ottenendo un notevole miglioramento in termini di costi temporali.

Il secondo vantaggio è quello di avere operazioni che sono più veloci dato il basso coefficiente di data rate dei video compressi e che possono sfruttare le caratteristiche precalcolate della codifica che si rivelano molto utili per il processo di segmentazione. Forniamo ora qualche indicazione sul funzionamento del flusso MPEG per poi passare a descrivere alcuni metodi usati per la sua

segmentazione temporale.

Rispetto al tipo di informazione usata questi metodi possono essere divisi in sei gruppi separati:

1) Coefficienti DCT .
2) Termini DC.
3) Termini DC, codifica a macro blocchi (MB) e motion vector.
4) Coefficienti DCT, codifica a macro blocchi (MB) e motion vector.
5) Codifica a macro blocchi (MB) e motion vector.
6) Codifica a macro blocchi (MB) e bitrate.

Non andremo a trattare tutti questi metodi ma descriveremo le due tecniche elementari che sono alla base di essi.

Il nome *MPEG* è l'abbreviazione per lo standard definito dalla *Moving Picture Expert Group* ed è attualmente lo standard più diffuso di compressione per i video digitali. Un'idea importante di questa codifica è quella della *compensazione di moto.* La maggior parte dei fotogrammi in una sequenza video sono molto simili tra loro, diventa quindi sensato pensare di non trasmettere le parti che rimangono fisse, e di evidenziare tramite un oggetto che ne indichi verso e direzione quelle che invece si sono spostate (*motion vector).* Seguendo questa idea si identificano tre tipi di fotogrammi:

- I fotogrammi che vengono codificati singolarmente senza alcun riferimento ad altri *(Intraframes o I frames)*
- I fotogrammi che vengono predetti basandosi su un fotogramma di tipo I *(Forward predicted frames o P frames)*
- I fotogrammi ottenuti interpolando fra un fotogramma I ed un fotogramma P *(Bidirectional frames o B frames)*.

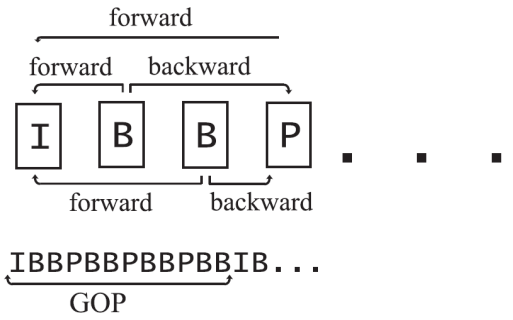

Figura 6: GOP e relazioni tra I, P e B

In tale maniera è possibile descrivere un video, attraverso una serie di fotogrammi legati tra loro da relazioni di successioni, quest'ordinamento viene chiamato *GOP (Group of Pictures),* una rappresentazione è data in Figura 6.

I fotogrammi di tipo *intra (I)* forniscono punti di accesso casuale tra i dati compressi e sono codificati utilizzando soltanto l'informazione presente nell'immagine stessa. Ogni blocco dell'immagine è trattato con una *Trasformata Discreta del Coseno (DCT).* I coefficienti *DCT* risultanti sono poi quantizzati e riordinati e poi codificati attraverso la *Run Length Encoding (RLE)* ed infine viene applicata uno schema a tavola fissa della codifica di Huffman (Figura 7).

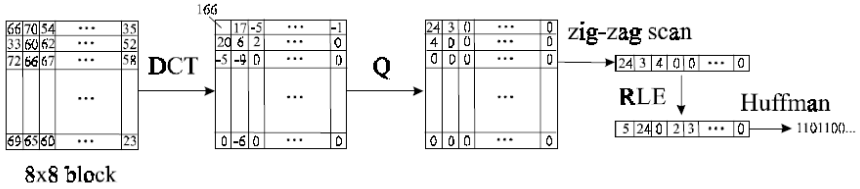

Figura 7: Codifica per fotogrammi di tipo Intra (I)

I fotogrammi *predetti (P)* sono codificati tramite una compensazione di moto in avanti *(forward)* utilizzando le precedenti immagini (I o P) più vicine.

Le immagini *bi-direzionali (B)* sono anch'esse compensate, questa volta in maniera consistente rispetto i fotogrammi passati e futuri.

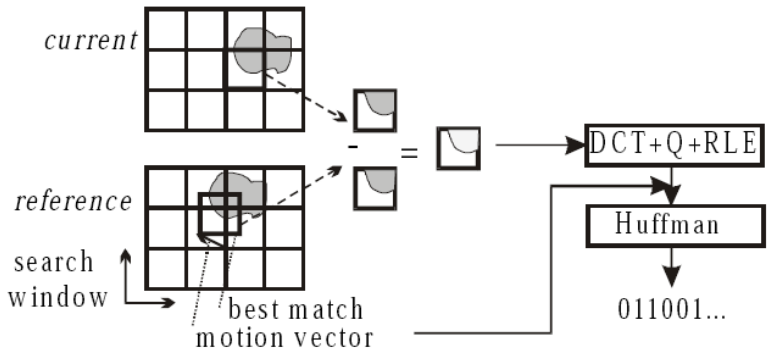

Figura 8: Predizione in avanti dei fotogrammi P

Per ciascun *macro blocco (MB)* di 16x16 pixel del fotogramma corrente, il codificatore cerca quello che deve diventare il miglior "campione di riferimento" per il fotogramma analizzato, cioè un'area nell'immagine precedente e successiva che è la più simile ad essa. Per far questo si codifica un vettore di movimento o *motion* vector utilizzato per descrivere le relazioni tra il macroblocco corrente e il campione di riferimento da cui è stato predetto. Di solito, dopo aver applicato la compensazione di moto, c'è ancora una differenza o un "residuo" tra il macro blocco ed il campione

di riferimento da considerare(Figura 8 e 9).

Durante il processo di codifica vengono quindi effettuate prove su ciascun macro blocco dei fotogrammi P e B per vedere se diventa più costoso usare la compensazione di moto o la codifica intra.

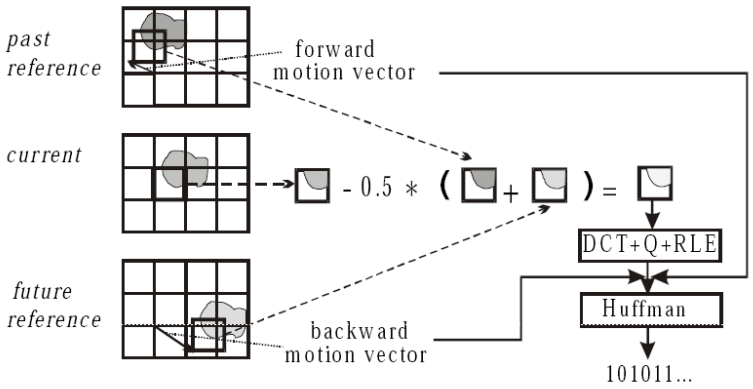

Figura 9: Predizione interpolata per i fotogrammi B

Queste codifiche e queste informazioni fornite dallo standard MPEG permettono quindi di attuare tecniche per la segmentazione temporale. Ad esempio nel caso dei fotogrammi *intra (I)* si possono andare ad analizzare i coefficienti DCT che forniscono una codifica esauriente del fotogramma in quanto:

1. Dato un sotto insieme di macro blocchi ed un sottoinsieme dei coefficienti DCT è possibile caratterizzare e riassumere il contenuto di un fotogramma rispetto ad altri.
2. La codifica permette di effettuare un'analisi tra i fotogrammi distanti *k* in maniera piuttosto veloce.

Si può quindi ad esempio rilevare una transizione se la correlazione tra i due fotogrammi è minore di una soglia prestabilita relativamente ai vettori dei coefficienti DCT:

$$D_{cut} = 1 - \frac{\vec{C}_f \cdot \vec{C}_{f+k}}{|\vec{C}_f||\vec{C}_{f+k}|}$$

Formula 9

Una seconda analisi è possibile andando a sfruttare le codifiche dei fotogrammi di tipo *B*. Se vi è discontinuità tra due fotogrammi consecutivi, ossia se c'è una transizione netta come uno shot, si può di solito associare un errore alto, corrispondente quindi a pochi *motion vector* nella codifica. Controllando quest'errore è possibile rilevare cambiamenti di immagini bruschi compiendo relativamente semplici calcoli sulle caratteristiche della codifica.

Queste due idee base ci permettono di costruire le varie tecniche di rilevamento delle transazioni, ad esempio sperimentalmente si è visto che utilizzare soltanto i fotogrammi di tipo I nel rilevamento delle transizioni tra shot risulta in un alto tasso di falsi positivi. Una soluzione ottimale prevede di combinare le due idee precedenti. Sfruttando sia i fotogrammi di tipo I che quelli di tipo B in una combinazione di analisi statistiche delle immagini prese dai fotogrammi I e dai vettori movimento associati con i fotogrammi B.

4 Analisi ed annotazione di un documento video

Al fine di poter estrarre contenuti da un flusso video lo si deve prima sezionare nei suoi elementi costituenti ed in seguito si devono fornire degli indici ai contenuti presenti in esso. Esistono diversi tipi di indici legati ai contenuti semantici. Ad esempio si vorranno estrarre informazioni relative agli oggetti presenti in un fotogramma chiave e confrontarle con quelle di un immagine fornita come interrogazione. Più difficile risulta l'estrazione di contenuti legati al movimento a causa della grande variabilità di casi presenti.

4.1 Analisi del contenuto di scena

Delimitare il campo di studio a precise tipologie di documenti video semplifica di molto il lavoro da fare, le principali tipologie sono:

- *Notiziari e telegiornali:* questi documenti presentano una struttura piuttosto definita e le immagini all'interno di questi presentano una funzione ausiliaria rispetto alle parole. La maggior quantità di dati semantici provengono in questo caso dalle parole di colui o colei che commenta il filmato, e queste spesso non sono strettamente correlate con la sequenza di immagini riprodotte. Esempi di indici sono il logo del telegiornale, il nome del cronista, il titolo della notizia.
- *Video sportivi:* dato il contesto specifico di questi documenti si può far uso di tecniche di analisi specializzate e mirate all'estrazione di parti rilevanti della sessione sportiva, o per la creazione di una visuale a più prospettive dell'azione di gioco. Esempi di indici sono l'estrazione del giocatore, il tracciamento del movimento del pallone o dei giocatori presenti su un campo di calcio.
- *Video commerciali:* l'intento è quello di fornire un impatto percettivo forte rispetto agli stessi contenuti. I colori scelti, le comparse e le tecniche di ripresa sono selezionate per costituire un messaggio pubblicitario ben definito. L'estrazione classica di oggetti all'interno delle immagini del filmato sono in questo caso di scarso interesse. Più rilevante è invece

l'estrazione del messaggio che viene comunicato. Inoltre il largo uso di effetti speciali e di contaminazioni tra grafica e cinematografia permette una precisa identificazione degli shot. Esempi di indici sono gli effetti di montaggio, il colore dominante nel movimento di un oggetto, la divisione spaziale tra il prodotto pubblicizzato ed il resto della scena.

- *Film:* il contenuto delle singoli immagini e il loro arrangiamento temporale diventa nei prodotti cinematografici essenziale per la comprensione della trama. Diventa quindi non solo importante l'analisi degli oggetti presenti nei fotogrammi ma soprattutto la correlazione tra essi. La produzione video, il montaggio di una sequenza di shot, il sonoro e la musica giocano tutti ruoli estremamente significativi.

Nei sistemi di video retrieval gli indici che vengono estratti sono normalmente espressi tramite parole chiavi testuali, o tramite un insieme strutturato di concetti, questa procedura è difficilmente realizzabile in maniera totalmente automatica e deve essere quindi accompagnata dall'intervento umano. In ogni caso l'estrazione automatica di questi indici tramite l'uso di algoritmi di analisi delle immagini e tramite regole appropriate è comunque possibile.

4.2 Analisi del movimento della telecamera

Analizzare i movimenti della videocamera in un filmato è utile per l'indicizzazione e per il recupero, questo perché rende possibile la segmentazione di lunghe sequenze in unità più corte ed omogenee definite dalla ripresa stessa. I movimenti della videocamera possono essere estratti sia dai flussi video compressi che da quelli non compressi.

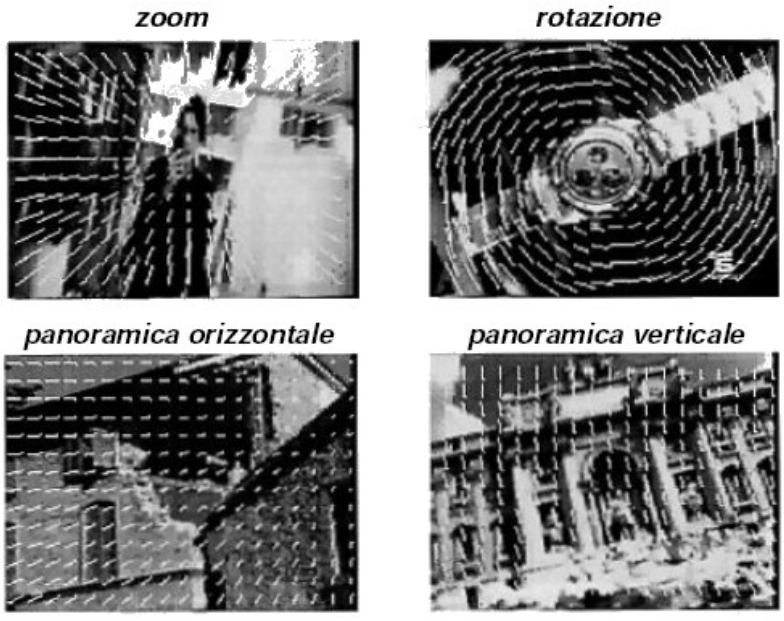

Figura 10: Quattro tipologie di movimenti di camera.

L'approccio classico a questo tipo di problema si basa sullo studio di un campo di flusso approssimante il campo di movimento dei pixel tra due fotogrammi consecutivi. Quando si analizzano flussi video compressi questo viene fatto attraverso i vettori di movimento. Questi vettori vengono utilizzati in accordo con l'algoritmo di confronto dei blocchi utilizzato nella compressione video dello standard MPEG. Come già spiegato i fotogrammi vengono partizionati in blocchi. I vettori movimento sono quindi quei vettori che minimizzano una funzione delle differenze tra le intensità dei pixel corrispondenti nel blocco. Solo i vettori con errori residui piccoli sulla compensazione di moto sono utilizzati per stimare il movimento. Esistono tre operazioni base della telecamera conosciute con la sigla *PTZ* queste sono la panoramica orizzontale(*pan*), la panoramica verticale (*tilt*) e quella in profondità (*zoom)*.

Nel panning e nel tilting, i vettori di movimento si diffondono sui fotogrammi con direzionalità precisa indicata da un *vettore modale*. Questi movimenti della camera vengono quindi rilevati tramite una soglia sulla differenza in valore assoluto tra i vettori movimento Θ_k e il vettore modale Θ:

$$\sum_{N}^{k} |\theta_k - \theta| \leq \Theta_p$$

Formula 10

L'operazione di zoom determina invece che i vettori movimento escano o entrino nel punto focale del fotogramma. Un'operazione di zoom viene rilevata dall'osservazione delle componenti verticali dei vettori di movimento. Durante uno zoom queste infatti possiedono nella prima e nell'ultima riga del fotogramma segni opposti, e quindi si controllerà che la loro differenza superi in grandezza entrambe le componenti.

4.3 Analisi del movimento degli oggetti

Un problema scientifico rilevante è dover distinguere tra il movimento reale di un oggetto all'interno della scena ed il movimento apparente indotto dalla telecamera. Esistono diversi approcci a questo problema basati sui cambiamenti di luminosità, sui vettori di movimento e sulle informazioni relative al colore.

Un modo è considerare il movimento simultaneo di più oggetti sulla scena, descrivendo ciascuno di essi tramite l'uso di parametri. Le scene statiche vengono descritte attraverso un modello di movimento singolo o di movimento dominante, ogni variazione da questo modello identifica degli oggetti che si muovono sulla scena.

Un'altra proposta è quella di estrarre degli *oggetti chiave* dagli shot ed interrogare il documento

video in relazione ai loro attributi visuali. Questi oggetti chiave non corrispondono agli oggetti semantici ma sono invece rappresentativi delle regioni che si muovono con movimento coerente. Per estrarre queste informazioni si usano tecniche di segmentazione del moto, e quando questo non è possibile si misura approssimativamente il moto tramite parametri statistici.

4.4 Analisi della colonna sonora

La comprensione di un documento video necessita dello studio non solo dell'immagine ma anche della componente sonora. Questa spesso è necessaria ed indispensabile all'analisi del contenuto di un documento multimediale e si richiede quindi che vada studiata in relazione alla componente video.

Lo studio dell'audio di un filmato tipicamente si compone di tre fasi:

- La rilevazione, segmentazione e rappresentazione della voce, della musica e dei suoni artificiali.
- L'identificazione del locutore nei sonori parlati.
- La trascrizione ed etichettatura del linguaggio naturale.

Lo studio del documento sonoro viene condotto analizzando l'energia contenuta nel segnale audio. L'unità elementare in cui esso viene diviso sono gli *audio frame*, corrispondenti a pochi millisecondi del segnale. I passi comuni per l'analisi audio sono principalmente tre:

- *Parametrizzazione dell'audio:* suddividere per finestre temporali il segnale audio ed estrarre caratteristiche di basso livello, normalmente nel dominio delle frequenze o sulle caratteristiche cepstrali.
- *Quantizzazione:* trasformare le caratteristiche di basso livello in qualche caratteristica statistica di alto livello.
- *Calcolo delle distanze:* calcolare le distanze tra le caratteristiche di alto livello dai differenti campioni audio, basandosi su determinati tipi di misurazioni.

Per distinguere tra musica e parlato esistono algoritmi che riescono a svolgere questo compito con una discreta accuratezza, sfruttando il fatto che parlato e musica hanno due distribuzioni spettrali e temporali diverse.

Il riconoscimento dell'audio all'interno di un video diventa essenziale in quella tipologia di filmati quali telegiornali, bollettini meteo o documentari, in cui è possibile assorbire più informazioni semantiche dalla componente sonora che da quella visiva.

4.5 Annotazione del video tramite icone

Un altro modo per fornire un indicizzazione ai contenuti presenti nel video e fornire un'annotazione visuale ad icone al documento, questo può essere fatto a mano. Si distingue questo processo in due rappresentazioni distinte:

1. Una *rappresentazione semantica*, che include le categorie degli oggetti, indipendentemente dall'ordinamento temporale delle loro azioni e della loro comparsa nella scena.
2. Una *rappresentazione temporale*, che crea relazioni specifiche lungo le categorie degli oggetti, attraverso la loro combinazione e l'ordinamento temporale delle loro azioni

Stabilite le due rappresentazioni bisogna solo procedere all'annotazione del video per fare questo si può procedere associando manualmente le icone a parti del video. Queste icone rappresentano visualmente le *categorie* (quali clima, condizioni di luce, locazioni, caratteri ed oggetti) oppure le *situazioni* (quali movimento di un carattere o di un oggetto) presenti nel video. L'utilità di una rappresentazione simile è la facilità portata dalla sua natura visuale e permette di fornire un tipo di annotazione efficiente.

Figura 11: Panoramica del sistema Silver

Un altro modo di annotare tramite icona è quello di sfruttare rappresentazioni ad esempi. Si costruiscono esempi visuali, che rappresentano approssimativamente la scena riprodotta nel contenuto di un segmento video al fine di ottenere una buona descrizione.

Implementazioni che seguono questo secondo approccio sono ad esempio il sistema *IMPACT*, e *Silver* della Carnegie Mellon University (Figura 11). Questi sistemi mirano ad enfatizzare la costruzione di una vista globale e strutturata del documento video, mirando ad una visione e

navigazione veloce dei contenuti, mostrando la linea temporale, le anteprime, una trascrizione del parlato e così via.

5 Visualizzazione dei fotogrammi chiave

In molti sistemi come quelli di video sorveglianza è utile fornire una visuale globale del contenuto del filmato. Un modo per fare questo è fornire una collezione di immagini salienti ridotte a piccole dimensioni proporzionali allo schermo e visualizzate in sequenza temporale. Il modo più semplice per farlo è quello di suddividere il flusso video in porzioni uguali e mostrare questi come "riassunto" di uno shot o di un filmato. Questi segmenti vengono poi ordinati e mostrati all'utente in maniera gerarchica (i cambiamenti di camera ed il contenuto video non sono presi in esame).

Una maniera più sofisticata è selezionare dei *fotogrammi chiave* da ciascuno shot, in maniera tale che i loro contenuti riflettano il contenuto semantico e le relative successioni ed evoluzioni nel video. La visualizzazione dei fotogrammi chiave ha il difetto che il numero di fotogrammi chiave estratto è di solito molto alto, specialmente per i film o i documenti. Questo può creare problemi se si vuole ottenere una visualizzazione globale e rende difficile catturare l'intero video in una singola vista.

Nell'utilizzazione di fotogrammi estratti per la rappresentazione di un video shot si presentano due problemi rilevanti che sono la *schematizzazione* e la *riduzione intelligente*.

Quando i professionisti dei media o gli analisti disegnano su carta una rappresentazione dei contenuti visuali di uno shot, evidenziano una vista semplificata delle caratteristiche salienti dell'immagine. Queste sono rappresentazioni visive molto efficienti, ci si chiede se non sia possibile realizzare automaticamente questi schemi. Sfortunatamente questo non è un problema di facile soluzione. Il secondo problema è legato all'esistenza di limiti sulla risoluzione alla quale i contenuti visuali di un immagine video sono leggibili. La maggior parte delle rappresentazioni nelle interfacce visuali usano risoluzioni che variano da 80x60 a 160x120 pixel. Occasionalmente quando si vuole elencare più immagini, la risoluzione scende a 40x30 pixel, ma così facendo molte immagini diventano illeggibili anche se rappresentano un buon punto di vista.

Questo può limitare la visualizzazione globale nella rappresentazione di un video, come soluzione si sono proposti metodi per la costruzione automatica di un'immagine rappresentante tutti i contenuti visibili in uno shot.

Questi metodi fanno uso di un'analisi dei movimenti di camera e di trasformazioni geometriche associate ad essi, le immagini successive vengono associate ad un fotogramma comune, e l'immagine di sintesi viene costruita passo passo. Questa immagine si presenta come una sintesi

dell'intera sequenza, tali immagini vengono chiamate *salient still*. Questi fotogrammi non rappresentano un momento finito di tempo, come può essere un singolo shot di un fotografo, ma aggregano i cambiamenti temporali che occorrono in una sequenza di immagini, fornendo le caratteristiche essenziali dell'immagine (Figura 12).

Figura 12: Sequenza di fotogrammi, l'immagine saliente selezionata è evidenziata in rosso.

Un'altra tecnica molto utilizzata è quella del *mosaico* in cui si riesce a rappresentare un intero shot tramite un fotogramma singolo nel quale vengono incluse tutte le informazioni statiche e dinamiche senza ridondanze. Un immagine mosaico è costruita da tutti i fotogrammi di uno shot allineando questi tramite un sistema di coordinate fisse e poi integrandoli (attraverso filtri temporali) in una singola immagine che fornisca una vista panoramica della scena. Gli oggetti in primo piano sono sovra imposti su un immagine di sfondo statica. Questa tecnica si basa sull'analisi dei parametri di movimenti, e da questi si derivano diversi livelli, ciascuno di essi rappresentante un singolo oggetto in movimento.

6 Formulazione dell'interrogazione

Le interrogazioni ad un sistema di Video Retrieval possono essere formulate in diverse maniere, le principali sono:

- Interrogazione per sketch visuali o per caratteristiche
- Interrogazione per esempio (*Query By Example*)
- Interrogazioni basate sul movimento
- Interrogazioni testuali

I sistemi visuali basati su disegni calcolano la correlazione tra lo sketch e la mappa dei bordi di ciascuna immagine nella base di dati, oppure utilizzando altre misure di distanza come le Wavelet. Alcuni sistemi, tra cui il sistema VideoQ da noi provato permettono di animare lo sketch stesso per dare la possibilità di specificare oltre a texture, forma e colore dell'oggetto d'esempio anche una sequenza temporale e di movimento come mostrato in Figura 12.

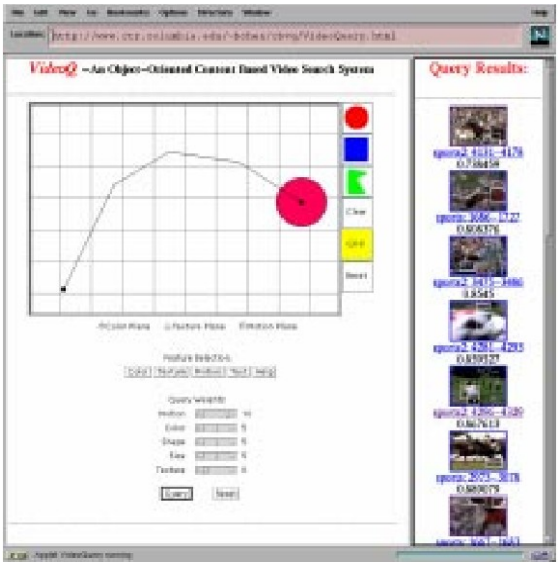

Figura 13: Sistema di sketch di VideoQ

I sistemi QBE funzionano con l'assunto che dato che il confronto corretto si avrà con un elemento all'interno del database, si può iniziare a cercare partendo appunto da un esempio preso da esso. Questa visione implica la speranza che si possa guidare l'utente attraverso una successione di interrogazioni fatte per esempi. Nel QBE si possono utilizzare schemi di partizionamento dello spazio per calcolare in anticipo gruppi gerarchici di immagini al fine di velocizzare la ricerca. Computando in anticipo le gerarchie di immagini sebbene la ricerca si velocizzi si ha una perdita dovuta alla natura dei raggruppamenti che in quanto statici necessitano di una computazione ogni volta che un nuovo video viene inserito nel database. Si vuole quindi che il QBE sia facilmente estendibile verso nuovi video, ottenendo un buon compresso tra i tempi di ricerca e quelli di aggiunta di nuovi dati.

Un altro approccio può essere vedere come elemento centrale di un interrogazione il movimento degli oggetti all'interno di esso. Il problema anche qui è come approcciarsi all'interrogazione, alcune idee sono quelle di poter rappresentare il movimento di forme semplici o di immagini fisse. Il sistema *VideoQ* permette ad esempio di far formulare all'utente tramite tool grafici una domanda del tipo: “Un rettangolo che si muove dritto verso destra” tramite la tecnica dello sketch.

L'ultimo approccio è quello testuale, che viene normalmente associato a quelli precedenti. Il permettere all'utente l'inserimento di interrogazioni sotto forma di testo rompe il limite di essere vincolati dagli strumenti grafici forniti. L'abbozzare un disegno può diventare molto complicato in quanto l'approccio visuale si presta a maggiori interpretazioni rispetto a quello testuale.

Infine la maggior parte dei sistemi di recupero supporta la formulazione di interrogazioni tramite descrizioni testuali. Per far questo si presuppone che i filmati che compongono la base di dati siano

stati precedentemente annotati tramite uno dei sistemi precedentemente descritti, e si presuppone che siano organizzati in maniera strutturata. Alcuni sistemi ad esempio forniscono dei linguaggi di interrogazione come il *videoSQL* supportato dal sistema OVID, in cui la clausola *Select* permette di selezionare qualsiasi tipo di oggetto video, e dove i vincoli vengono specificati sugli attributi all'interno di una clausola *Where*, ad esempio con un interrogazione simile:

```
SELECT  vid:[s,e]
FROM video:VidLib
WHERE (vid,s,e) IN VideoWithObject(Dennis) AND
object IN ObjectsInVideo(vid,s,e) AND
object != Dennis AND
      typeof(object) = Person
```

Importante notare che non esiste un sistema di interrogazione ed un interfaccia unica per ogni sistema di video retrieval, ma bisogna scegliere una combinazione opportuna delle tecniche precedentemente elencate alla fine di riuscire ad incontrare la visione personale dell'utente che interroga il sistema.

7 Conclusioni

In questa breve relazione abbiamo cercato di fornire un introduzione alle tecniche di Content Based Video Retrieval cercando di mostrare la varietà di queste tecniche e cercando un approccio verso i concetti e alle idee base che vengono applicate per risolvere i problemi in questo campo di studi. Nel periodo di stesura della relazione abbiamo anche provato sperimentalmente alcuni sistemi, cercando dove possibile applicazioni software di libero accesso; purtroppo a parte il sistema *VideoQ* abbiamo riscontrato che ci sono poche applicazioni pratiche di questi sistemi e quelle che ci sono sono proprietarie e a pagamento. Un altro problema che abbiamo dovuto affrontare è stata la mancanza di documentazione chiara sull'argomento, questo per la natura profondamente innovativa e di ricerca di queste tecnologie. Riteniamo utile comunque questo lavoro perché ci ha permesso di approfondire e di riunire diversi argomenti visti a lezioni relativi al recupero di informazioni sia dai documenti visuali che da quelli sonori.

Riferimenti

1. Alberto Del Bimbo, *Visual Information Retrieval*, Morgan Kauffman, (1999).
2. P. Aigrain, HJ Zhang, D. Petkovic, *Content-based Representation and Retrieval of Visual Media: A state-of-the-Art Review*, Multimedia Tools and Applications, (1996).
3. I. Koprinska, S. Carrato, *Temporal Video Segmentation: A Survey*, (2001).
4. Richard Hallows, *Tecniques used in the content-based retrieval of digital video*, University of Southampton, (2001).
5. J.S. Boreczky and Lynn D. Wilcox, *A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features*, X Palo Alto Laboratory Palo Alto, (1998).
6. S.Chang, W.Chen, H.J. Meng, H. Sundaram, D. Zhong, *VideoQ: An Automated Content Based Video Search System Using Visual Cues*, ACM Multimedia, (1997).
7. A. Akutzu, Y. Tonomura, *Video Tomography: An Efficient method for Camerawork Extraction and Motion Analysis,* ACM Multimedia (1994).
8. J. Casares, B.A. Myers, A.C. Long, R. Bhatnagar, S.M. Stevens, L. Dabbish, A. Corbett, *Symplifying Video Editing with Intelligent Interaction,* Carnegie Mellon University, (2001).