Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Estrazione e Descrizione MPEG-7 di Oggetti Video

M. Tripiciano[1], M. Trapani[2],

RT-ICAR-PA-03-14 **dicembre 2003**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) –
Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: www.icar.cnr.it
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: www.na.icar.cnr.it
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: www.pa.icar.cnr.it

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Estrazione e Descrizione MPEG-7 di Oggetti Video

M. Tripiciano[1], M. Trapani[2],

Rapporto Tecnico N.: **RT-ICAR-PA-03-14**

Data: **dicembre 2003**

---

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sezione di Palermo .
[2] Tesista Università degli Studi di Palermo Dipartimento di Ingegneria Informatica.

Video Object Extraction and MPEG-7 metadata description

M. Tripiciano, M. Trapani

---

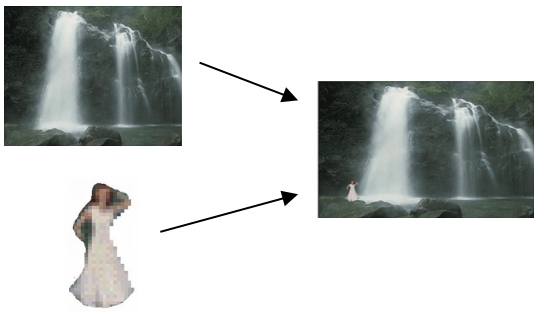
1 Introduzione

Negli ultimi anni è nato un nuovo modo di considerare un filmato video, non più come una semplice sequenza temporale di fotogrammi, bensì come una combinazione strutturata di più oggetti video che ad esso appartengono; un Video Oggetto può, quindi, essere una semplice sequenza temporale di frames (video object temporale), può identificare una o più regioni di uno o più frames (video object spaziale) o, infine, essere l'insieme di entrambi le accezioni individuando un oggetto in movimento (video object spazio-temporale).

Questa interpretazione di un filmato è nata poiché una tale organizzazione permette una maggiore flessibilità; ad esempio si potrebbe visualizzare solo un oggetto, combinare oggetti provenienti da filmati diversi oppure inserire informazioni aggiuntive ed elementi non reali. Con una opportuna descrizione di tutti gli oggetti che compongono il video è possibile organizzare dei databases di video object per consentire un facile reperimento e riutilizzo dell'informazione desiderata.

Lo scopo di questo rapporto è lo studio e lo sviluppo di alcune metodologie che permettano ad un utente, nella maniera più semplice possibile, l'individuazione inseguimento ed estrazione degli oggetti di interesse di un filmato , il calcolo di alcune features di basso livello (colore, movimento) ed il relativo salvataggio dei dati e dei metadati che descrivono le caratteristiche di tali oggetti.

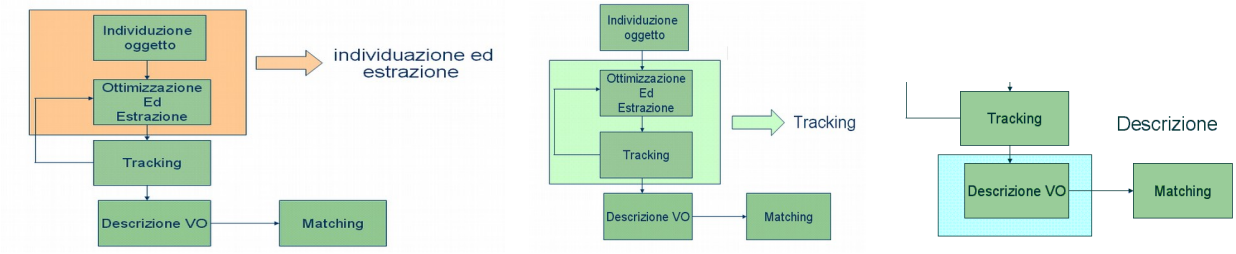
Il criterio da utilizzare per la scelta degli oggetti da descrivere dipende dall'applicazione e dal tipo di riutilizzo che di tali oggetti si intende fare; i campi di applicazione possono essere i più svariati dall'editing video a sistemi semi-automatici per il processamento dell'informazione (video restauro).
Il video oggetto è un entità di un filmato che una volta estratto può esistere anche indipendentemente dal filmato originario; tale indipendenza permette di modificare o aggiungere particolari per comporre nuovi filmati (video editing)



Uno dei possibili utilizzi potrebbe essere la composizione di video a scopo didattico o divulgativo. Tramite la combinazione di elementi reali ed elementi descrittivi si può comporre un video in cui l'utente in modo interattivo selezionando un oggetto potrebbe avviare la visione di un particolare evento o personaggio oppure avere delle informazioni supplementari riguardanti l'oggetto selezionato.

2 Video Objects indexing

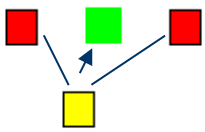
Allo scopo di trattare i video objects significativi del filmato è necessario definire e sviluppare alcuni moduli software ciascuno dei quali si occupa di una delle fasi della catena logica-temporale di elaborazione: individuazione, estrazione ed inseguimento ed, infine, descrizione dei metadati relativi alle caratteristiche visuali degli oggetti individuati.
Nel diagramma seguente vengono mostrate le diverse fasi della catena di elaborazione.



Nella prima fase il modulo di individuazione è caratterizzato dall'interazione dell'utente che deve individuare grossolanamente l'oggetto da estrarre dal filmato selezionandone alcuni punti del contorno; questa fase si può suddividere in diversi passi:

- selezione su display di un fotogramma contenente l'oggetto da definire
- selezione interattiva di alcuni punti del contorno
- inizializzazione del contorno (centro-gravità,distanza media punti…)
- ottimizzazione tramite funzioni di energia (snakes)

I contorni deformabili o snake chiusi o aperti, a seconda che il primo e l'ultimo punto siano collegati o meno, nascono come soluzione al problema riguardante l'individuazione dei contorni. Lo snake è un insieme ordinato di punti , $V = [v_1,v_2….,v_n]$ dove ogni punto v_i è individuato da una coppia di coordinate intere che individuano la sua posizione (x,y). Vengono calcolate due funzioni di energia, una esterna E_{ext} ed una interna E_{int}.
L'energia esterna ha come proposito principale quello di rendere il contorno più omogeneo possibile eliminando le asperità e quindi in assenza di una forza esterna cerca di rendere il contorno simile ad una retta nel caso di contorno aperto e simile ad un cerchio nel caso del contorno chiuso; il valore di tale energia viene calcolato come differenza della posizione del punto rispetto i due vicini e la posizione ottimale che dovrebbe avere.
L'energia interna cerca invece di spostare il punto verso il contorno; si ottiene, quindi, una mappa dei contorni con informazioni relative all'intensità ed alla direzione del contorno.



Come tirato da un elastico il punto giallo viene spinto verso il punto ottimale - punto verde (zero energia) . Limita le asperità nel contorno.

Generata una mappa delle forze dove un punto del contorno ha intensità maggiore, l'elemento dello snake ne viene attratto.

Lo snake in definitiva trasforma il problema dell'estrazione dei contorni in un problema di minimizzazione di energia. Quello che si ricerca è l'insieme dei punti V_o che minimizzano l'espressione dell'energia:

$$E_{tot} = \sum \lambda_i * E_{int}(\mathbf{v}_i) + (1-\lambda_i) E_{ext}(\mathbf{v}_i)$$

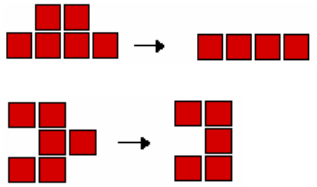
essendo λ_i il parametro di regolarizzazione che influenza significativamente la soluzione ottimale. Per ottimizzare la ricerca dell'energia totale viene applicato il principio del MinMax secondo il quale, per una situazione analoga a quella qui presente, un buon compromesso si ottiene minimizzando i valori dell'energia avendo scelto come λ il valore massimo. Tramite la ricerca tra le possibili combinazioni di punti che rendono minima l'espressione dell'energia, otteniamo il contorno ottimale. Tale ricerca utilizza una struttura piramidale dove ogni livello ha dimensione metà del livello inferiore.

Inizialmente i punti, coerentemente modificati, vengono ricercati in un livello della piramide più alto; terminata la ricerca il contorno viene espanso per mantenere una corretta equidistanza dei punti e passato al livello inferiore dove viene ricalcolato. Per ottenere un buon risultato è fondamentale la scelta del contorno iniziale; ecco perché è auspicabile che sia l'utente ad individuare interattivamente i punti iniziali.

Nella fase successiva di estrazione e tracking , il relativo modulo si propone ,avendo a disposizione il contorno ottimizzato nel primo fotogramma, di estrarre la maschera corrispondente all'oggetto ed inseguirlo ed estrarlo nei fotogrammi successivi e precedenti dando origine all'insieme di maschere che caratterizzano il video oggetto.
L'estrazione e costituita dai seguenti passi :

- Completamento del contorno.
  La lista di punti ottenuta è costituita da punti non contigui e dunque si rende necessario un completamento di tale contorno in modo tale che ogni punto abbia negli otto-vicini almeno due elementi del contorno. A questo punto potrebbero essere comparsi dei punti anomali che comprometterebbero l'estrazione dell'oggetto.

- "Pulizia" del contorno ottenuto.
  Vengono dunque effettuate le operazioni di thinning per "pulire" il contorno da punti non necessari stando attenti comunque a mantenere la continuità del contorno e di prouning per eliminare inutili code.



Thinning : ridimensionamento di parte del contorno

Prouning : "taglio" di sporgenze non necessarie

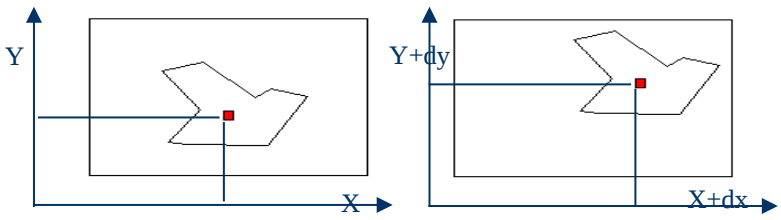
- Estrazione dei punti interni.
  Ripulito il contorno, tramite una unica scansione, vengono individuati i punti appartenenti all'interno del contorno e viene dunque generata la maschera binaria.

Individuato l'oggetto di interesse nel primo fotogramma questo viene inseguito nei fotogrammi successivi. Naturalmente per riuscire ad ottenere dei buoni risultati si dovrà avere una previsione del movimento dell'oggetto e per fare ciò ci si può servire delle mappe di moto circular zone search (CZS); tali mappe danno delle informazioni sul movimento di piccoli blocchi in cui viene suddivisa l'immagine, indicando la velocità e la direzione dello spostamento.

L'approccio è quello di vedere la posizione del centro di gravità dell'oggetto nel primo fotogramma e ricalcolare la sua posizione nei fotogrammi successivi ottenendo, dunque, per ogni punto un valore di Dx , Dy e di affidabilità.

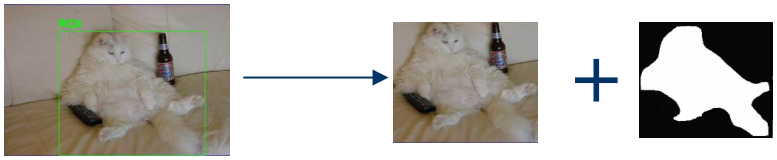
I punti appartenenti al contorno sono quelli con minore affidabilità poiché essendo l'oggetto in movimento difficilmente un blocchettino contenente parte di sfondo rimarrà immutato. Per tale motivo si è scelto di calcolare lo spostamento tramite punti interni all'oggetto.

Ottenute le informazioni relative alla posizione dell'oggetto nel fotogramma e posizionato lo snake, grazie alla sua capacità di riadattamento, saremo in grado di ottenere il contorno desiderato.



3 L’organizzazione dei dati

Estratto l’oggetto dal filmato i dati vengono memorizzati in un server di video oggetti. Le dimensioni di ogni fotogramma sono quelle della minima regione di interesse (ROI) che contiene l’oggetto. Per ogni frame vengono salvate una o più maschere relative a tinta, saturazione, luminosità (immagine a colori) e una maschera binaria che individua i punti che appartengono all’oggetto.



Un insieme di caratteristiche di natura intrinseca dell’oggetto quali il colore, la forma , la tessitura vengono calcolate in maniera automatica e permettono di associare al video object la sua descrizione in metadati e senza la quale non sarebbe possibile l’effettivo riutilizzo di un video oggetto; tale descrizione potrà ulteriormente essere arricchita tramite annotazioni con informazioni non ricavabili automaticamente ma ugualmente caratteristiche del VO.

La scelta delle features da estrarre od annotare per ogni oggetto o singolo frame deriva dall’utilizzo che è stato prefissato. Un primo utilizzo è quello del matching tra video oggetti appartenenti allo stesso filmato per valutare se lo stesso elemento si trova in fotogrammi non continui (es. cambi di inquadratura) per un successivo riassemblaggio. Un altro utilizzo potrà essere la ricerca per similitudine o descrizione in un database di video oggetti.

L’elenco che segue mostra le features descrittive principali riscontrabili a livello di meta-informazioni a livello di video, di oggetto e di singolo fotogramma o porzione di esso.

Film features	Oggetto features	Fotogramma features
Titolo del film	Nome oggetto	Locazione fisica
Durata	Fotogramma iniziale	Dimensione della ROI
Locazione	Durata	Origine della ROI
Dimensioni	Keyframe	Contorno dell’oggetto
Formato	Colore dominante	Coordinate centroide
Profondità colore	Tessitura	Velocità

Caratteristiche quali il titolo o informazioni come il luogo degli avvenimenti sono evidentemente delle annotazioni testuali, mentre altre caratteristiche quali colore, tessitura, contorno hanno una struttura descrittiva più complessa.
Per descrivere la caratteristica del colore,ad esempio, è stato scelto un metodo semplice e compatto che permette di confrontare due oggetti considerando il loro colore dominante individuato in un set di colori predefiniti (Max 8).

Il descrittore è definito tramite una terna di valori che rappresentano il colore dominante, la percentuale in cui è presente nell’oggetto, la coerenza spaziale; quest’ultimo valore rappresenta la differenza tra una grande macchia di colore omogeneo rispetto al colore sparpagliato in tutta l’immagine.

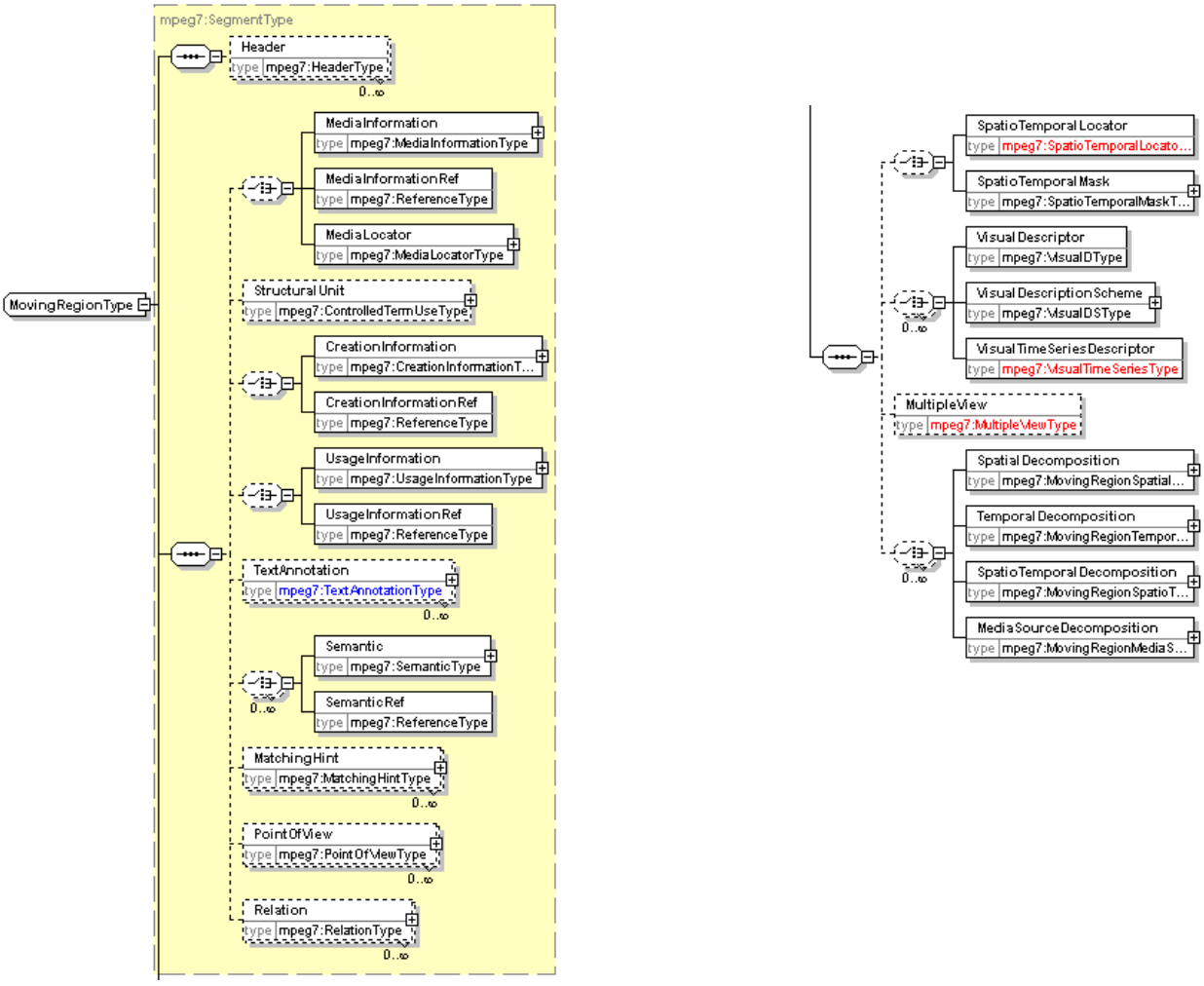
Per la tessitura è stato scelto l’Homogeneous texture descriptor (HTD) che fornisce una rappresentazione quantitativa della tessitura. Per il contorno vengono salvati i punti dello snake (ottimizzati) riferite all’origine della ROI e la posizione del centroide.

4 MPEG-7 Video Objects Description

Per descrivere sia le caratteristiche di carattere generale del video che quelle più specifiche riguardanti i video objects (VO), è stato scelto di utilizzare lo standard MPEG-7 (*Moving Pictures Expert Group*). Le motivazioni di tale scelta sono varie tra le quali quella di essere uno standard consolidato per la descrizione di documenti multimediali ed, inoltre, essendo basato sul formato di descrizione standard XML (*EXtensible Markup Language*)
, ne eredita tutte le peculiarità come la trasportabilità multipiattaforma, l'estendibilità e l'utilizzo di schemi descrittivi che specificano il tipo dei dati descritti.
Nello standard MPEG-7 sono stati definiti un vasto numero di descrittori semplici (D) e di schemi descrittori (DS) in modo tale da potere descrivere un documento multimediale sotto vari aspetti, da quello di carattere generale (titolo, autore, locazione…) a quelli di segmentazione, sommarizzazione, contenuto, semantico e non, agli aspetti di fruibilità da parte di potenziali utenti (profili utente, scenari…). In questo lavoro la nostra attenzione si focalizza sugli aspetti di descrizione degli oggetti significativi contenuti nel video ed, in particolare, le proprietà caratteristiche visuali di basso livello come colore e tessitura.

4.1 Schemi Descrittivi

I video object considerati hanno una loro evoluzione spazio-temporale e per la loro descrizione abbiamo scelto di utilizzare uno degli schemi descrittivi base di MPEG-7; il tipo Regione in Movimento (*MovingRegion*). L'oggetto individuato si muove all'interno di una regione per un intervallo di tempo T; lo schema descrittivo scelto permette di definire all'interno della sua struttura per ogni oggetto individuato, sia le informazioni relative alla locazione e movimento di tale regione (*SpatioTemporalLocatorType* DS), sia i descrittori visuali dell'oggetto (*DominantColorType* DS …). Di seguito è mostrato il diagramma della struttura gerarchica utilizzata dove le zone tratteggiate indicano rami dell'albero o descrittori opzionali, il simbolo  indica una sequenza ordinata degli elementi in esso definiti ed, infine, il simbolo  indica la scelta tra uno degli elementi definiti.



4.2 Features di basso livello (Visual Descriptors)

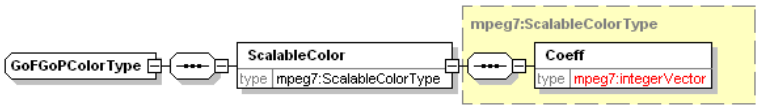
4.2.1 GoFGoP Color

Lo schema descrittore *GoFGoPColor* (Group of Frame Group of Picture Color) fornisce l'informazione riguardo la distribuzione di colore per l'intero intervallo temporale cioè, nel nostro caso dell'oggetto. In pratica viene calcolato il valore medio, mediano o di intersezione tra gli istogrammi di colore di tutti i frames o porzioni di esso nell'intervallo; avremo, quindi, nei tre casi, i tre rispettivi modi seguenti di calcolare di calcolare i valori del vettore risultante (*coeff*) da inserire nel descrittore *ScalableColor*:

$$Avg\_Histogram\_value[j] = \frac{1}{N}\sum_{i=0}^{N-1} Histogram\_value_i[j];\ j = 0,...,255.$$

$$Med\_Histogram\_value[j] = \mathbf{median}(Histogram\_value_0[j],...,Histogram\_value_{N-1}[j]),\ j = 0,...,255.$$

$$Int\_Histogram\_value[j] = \mathbf{min}_i(Histogram\_value_i[j]),\ j = 0,...,255.$$



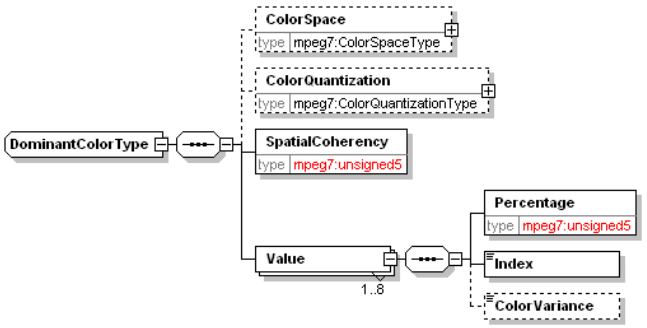
4.2.2 Colore Dominante

Questo schema descrittore (DS) specifica un set di colori maggiormente presenti in una regione di forma qualsiasi. Il suo scopo è quello di permettere una ricerca basata sul contenuto (colore) di immagini di forma arbitraria. Per ogni oggetto viene calcolata la feature "colore dominante" del suo keyframe.

Il descrittore *ColorQuantization* definisce la quantizzazione dello spazio di colore definendo dunque una mappatura dei componenti dei colori in un range di interi [0,NumOfBin-1]. Vengono specificati il numero di componenti del colore e per ogni componente il numero dei livelli.

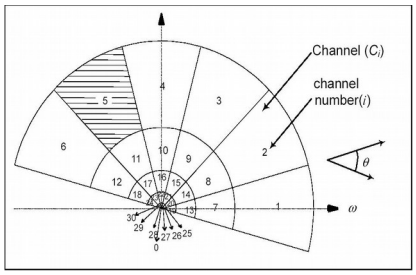
SpatialCoherency indica la coerenza spaziale dei colori dominanti ed è calcolato come la somma pesata della coerenza spaziale dei singoli colori dominanti. Il peso risulta essere legato al numero di pixel di ogni colore; esso rappresenta quanto contigui sono i pixel simili a secondo che si tratti di una macchia di colore o pixel sparsi nell'immagine.

Inoltre per ogni colore dominate avremo altri due descrittori:il primo, *Percentage* che specifica la percentuale di pixels che sono associati al colore in esame; tale valore è quantizzato uniformemente nell'intervallo 0, corrispondente allo 0%, e 31 , corrispondente al 100%.
Un secondo descrittore *Index* specifica il colore dominante nello spazio definito nel descrittore *ColorQuantization*.



4.2.3 Texture

L'*Homogeneous Texture* è un descrittore che caratterizza la tessitura in una regione utilizzando i valori dell'energia e della deviazione dell'energia (opzionale) in un set di canali di frequenza; tale descrittore si basa sul HSV (Human Visual System) che presuppone una suddivisione del dominio della frequenza in bande di ottave rispetto la direzione radiale ed in angoli di costante ampiezza lungo quella angolare. Avremo 5 bande ed angoli di 30° per un totale di 30 canali (C_i).



La frequenza ϖ ,normalizzata tra 0 ed 1, è data da $\varpi=\Omega/\Omega_{max}$. Essendo Ω_{max} il massimo valore di frequenza dell'immagine.

Ogni canale è caratterizzato dall'angolo θ_r
Individuato dalla relazione $\theta r=30°*r$ con $r\in\{0,1,2,3,4,5\}$

Average è un descrittore che specifica l'intensità media dei pixels dell'immagine. Tale valore viene quantizzato con 8 bits assegnando a 0 il valore minimo ed a 255 il valore massimo.

StandardDeviation è un descrittore che rappresenta la deviazione standard dell'intensità dei pixels dell'immagine. Anch'esso è quantizzato ad 8 bits ma con valore minimo uguale a 1.31 e massimo 109.48.

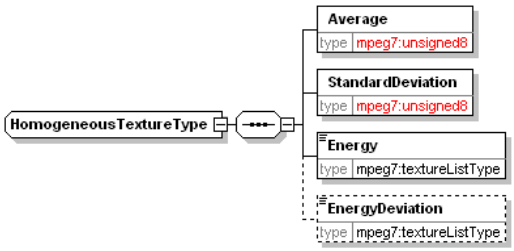
Per ognuno dei 30 canali viene, inoltre, calcolato il descrittore *Energy*. Si tratta di un vettore di trenta elementi dove ogni elemento rappresenta il valore di energia del rispettivo canale.

L'energia del generico canale ,calcolata nel dominio della frequenza e in relazione alle funzioni di Gabor, è definita come il logaritmo della somma al quadrato dei coefficienti della trasformata di Fourier filtrati dalle funzioni di Gabor.

$$e_i = \log_{10}[1+p_i],$$

dove

$$p_i = \int_{\varpi=0^+}^{1} \int_{\theta=(0^0)^+}^{360°} [G_{P_{s,r}}(\varpi,\theta) \cdot P(\varpi,\theta)]^2$$



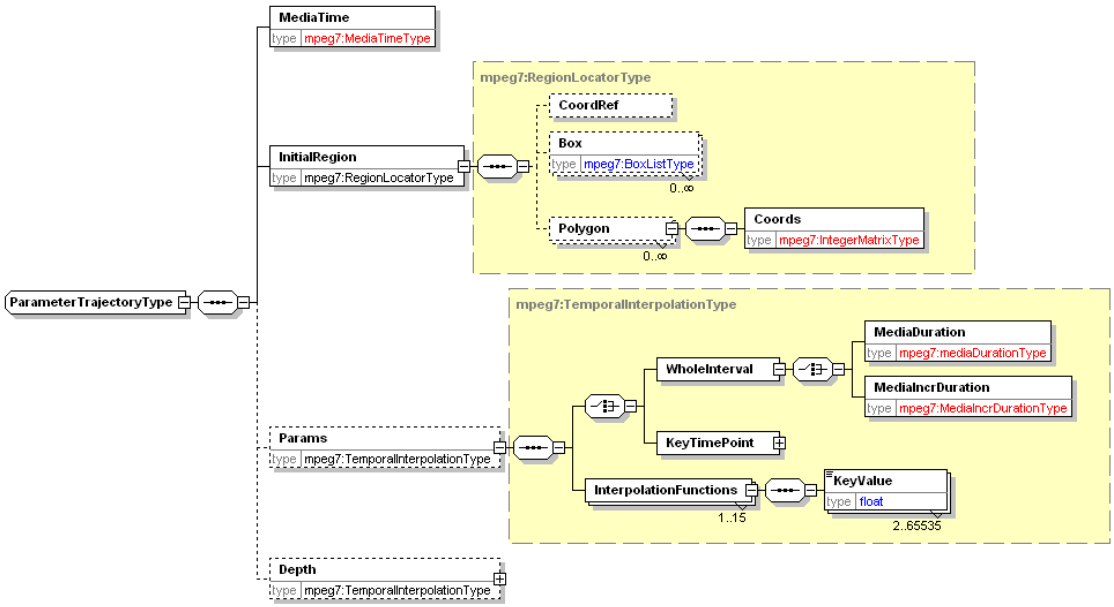
4.2.4 Regione e Contorno dell'oggetto

In questa sezione della descrizione troviamo i descrittori relativi alla definizione e posizione del box che circoscrive l'oggetto ed i parametri relativi al suo movimento nell'intervallo di tempo. Vengono descritti i parametri relativi alla sotto-regione iniziale e quelli relativi al sua traslazione nel tempo.

L'elemento *MediaTime* specifica il tempo iniziale e la durata della regione spatio-temporale descritta.

L'elemento *InitialRegion* specifica una regione di riferimento e la relativa struttura.

L'elemento *Params* specifica la traiettoria di un punto di riferimento della regione usando un modello parametrico di movimento.



5. References

[1] Kok Fug Lai "Deformable Contours: modeling,extraction,detection and classification." (University of Winsconsin-Madison 1995).

[2] S.Delgrado Olabarriaga "Human-Computer Interaction for the Segmentation of Medical Images"

[3] Hao Jiang and Mark S. Drew "A predictive contour inertia snake model for general video tracking" (School of Computing Science, Simon Fraser University,Vancouver, B.C., Canada)

[4] S. Lefevre, J.P. Gerard , A. Piron , N. Vincent "An Extended Snake Model For Real-Time Multiple Object Tracking" (Université de Tours- FRANCE -September 2002)

[5] World Wide Web Consortium's (W3C) XML web site http://www.w3.org/XML.

[6] Information Technology — Multimedia Content Description Interface – Part 3: Visual (ISO/IEC 15938-3:2002/FPDAmd 1)

[7] Information Technology — Multimedia Content Description Interface – Part 5: Multimedia Description Schemes (ISO/IEC 15938-5:2003/FDAmd 1)