

ToCAI: uno schema di descrizione per l'indicizzazione ed il recupero di documenti multimediali

A. Bugatti, R. Leonardi, P. Migliorati, e L. A. Rossi

Università degli Studi di Brescia, Dip. di Elettronica per l'Automazione

{adami, leon, pier, lrossi}@ing.unibs.it

Sommario

In questo articolo viene presentato uno schema di descrizione DS (Description Scheme) denominato con l'acronimo ToCAI (Table of Content-Analytical Index). L'idea di fondo è di sfruttare una modalità simile a quella adottata per capire l'organizzazione ed il contenuto di un libro tecnico per esplorare un documento Audio-Video (AV). Dall'indice ToC (Table of Contents) è possibile capire l'organizzazione sequenziale in cui sono ordinati i diversi capitoli mentre tramite l'indice analitico AI (Analytical Index) è possibile arrivare in modo veloce ad argomenti o termini di interesse. L'utilizzo di questo tipo di schema permette una rappresentazione gerarchica dell'organizzazione temporale di un documento AV, (tramite la sezione ToC) molto utile per esplorare in modo veloce, assieme ad un indice analitico (AI) che punta ad oggetti multimediali di interesse. Vengono esaminati anche due sotto schemi di descrizione corrispondenti rispettivamente a come è stata generata la descrizione ed ai metadata associati al documento. Viene descritta la struttura dettagliata del DS utilizzando la notazioneUML per passare poi ad un esempio reale. Infine sono state riportate alcune considerazioni sull'interfaccia realizzata.

1. Introduzione

Allo stato attuale è in costante crescita la disponibilità di materiale AV in formato digitale. Si rende pertanto necessario avere a disposizione degli strumenti che permettano ad un utilizzatore di muoversi in questa mole di informazioni in modo da riuscire a selezionarne una sottoparte di interesse. A questo proposito negli ultimi anni sono stati prodotti diversi contributi riguardanti l'indicizzazione e la ricerca di materiale AV tra cui [3],[8],[11]. L'International Standard Organization (ISO) ha dato inizio, nell'ottobre 1996, ad un processo per la standardizzazione del formalismo da utilizzare per descrivere il contenuto di documenti multimediali, chiamato MPEG-7: "Multimedia Content Description Interface" [6],[7]. Per settembre 2001 è auspicata la definizione di un insieme standard di descrittori (D) e di schemi di descrizione (DS) rappresentati tramite un apposito linguaggio chiamato "Description Definition Language" (DDL). Un DS si può utilizzare come base per istanziare una descrizione a più livelli di astrazione, di un documento multimediale, combinando descrittori associati a caratteristiche quali ad esempio forma, colore, tessitura, moto (per gli oggetti video) e/o caratteristiche audio (relative alla componente audio) [1]. Più in dettaglio il

DDL dovrà permettere la costruzione di diversi schemi di descrizione utili in contesti applicativi specifici.

In uno schema di descrizione che rappresenta le informazioni multimediali in maniera gerarchica, si può avere al livello di astrazione più basso descrizioni del contenuto che si limitano alla forma, dimensione, colore e tessitura degli oggetti che compongono ciascun fotogramma [5]. Ad un livello di astrazione più alto la descrizione del contenuto può essere data l'analisi del moto e della deformazione degli oggetti. Un ulteriore livello si può determinare in base agli stacchi di telecamera (shot) congiuntamente alla tipologia di effetti di editing utilizzati [2]. Se il processo di astrazione continua, la descrizione diventa progressivamente più significativa, ad esempio identificando le scene. Queste possono essere interpretate da adeguati modelli di conoscenza nel caso in cui sia presente una descrizione di tipo simbolico (ad esempio scena dove: un cane abbaia, una palla scorre verso destra, si sente il rumore di un motore). Per raggiungere un buon livello di descrizione dell'informazione è necessario considerare con la stessa importanza i segnali associati all'audio e al video. Sfortunatamente nel passato le ricerche su tecniche di analisi audio e video si sono sviluppate in modo indipendente. Solo recentemente queste due sorgenti di informazione sono state prese in considerazione in modo congiunto [9],[10]. Dai primi risultati è emerso che l'analisi congiunta dei due segnali è molto efficace per la classificazione/identificazione delle scene [10].

Per esempio si consideri una procedura in grado di identificare i momenti salienti di una partita di calcio. Un algoritmo che analizza il video per ricercare i goal si può combinare con uno audio che ricerca rapidi incrementi nell'energia del segnale. In sintesi si può dire che la descrizione del contenuto di materiale AV è di notevole attualità. Inoltre è plausibile ritenere che la migliore descrizione si possa ottenere utilizzando dove è possibile un'analisi congiunta dell'audio e del video. In questo lavoro viene proposto uno schema di descrizione che da una rappresentazione gerarchica del contenuto di un documento multimediale, considerando sia l'analisi video che quella audio.

Le funzionalità racchiuse in questo DS sono:

- Caratterizzazione della struttura temporale di un documento multimediale da un punto di vista semantico a più livelli di astrazione. Ogni livello è composto da una serie di segmenti consecutivi coerenti rispetto alla semantica associata
- Permettere di recuperare in modo semplice ed efficiente informazioni di interesse come oggetti particolari che compaiono in un video (per esempio Bill Clinton), o identificare specifiche situazioni di interesse (un omicidio in un thriller). Per avere delle buone capacità di ricerca, è importante che oggetti ed eventi siano disposti nell'AI secondo diversi criteri, tali da facilitare il lavoro di ricerca.
- Offrire informazioni generiche e specifiche di un documento multimediale come ad esempio autore, titolo, data di produzione, etc.
- Fornire informazioni utili relativamente alla descrizione stessa come, la sua dimensione, il tipo di tecniche di estrazione utilizzate e la loro affidabilità.

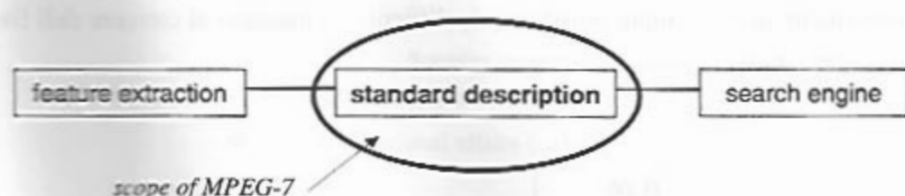


Fig. 1 L'obiettivo di MPEG-7.

L'idea originale che sta alla base di questo schema di descrizione si rifà alla struttura utilizzata nei libri tecnici. Si può capire in modo semplice l'organizzazione sequenziale di un libro guardando l'indice (che generalmente si trova nelle prime pagine) e/o ricercare elementi di interesse utilizzando l'indice analitico (tipicamente posto in fondo al libro). Nel primo caso viene conservato l'ordine cronologico degli argomenti presentati, mentre nell'ultimo viene utilizzato il criterio di ordinamento alfabetico per facilitare la ricerca. Il ToCAI presenta un meccanismo simile associato però ad un documento multimediale. In più però permette di recuperare informazioni a tutti i livelli di astrazione disponibili infatti in un libro una parola chiave dell'AI punta alla pagina che contiene quell'argomento e non al paragrafo o alla sezione corrispondenti.

Il presente articolo è organizzato nel seguente modo. Nella Sezione 2 vengono dati alcuni semplici concetti relativi ad MPEG-7 [6]. Nella 3 viene introdotta la struttura dello schema di descrizione ToCAI presentando le funzionalità dei sotto schemi che impiega. La sezione 4 tratta i dettagli del ToCAI e dei DS e descrittori che utilizza attraverso la notazione UML. Nella 5 viene presentato un esempio di implementazione del ToCAI mentre nella 6 sono riportati alcuni dettagli tecnici relativi all'interfaccia video sviluppata sulla base del DS qui proposto.

2. MPEG-7: contesto ed obiettivi

Nell'ottobre 1996, MPEG ha dato inizio ad un nuovo progetto riguardante la descrizione del contenuto di materiale multimediale: "Multimedia Content Description Interface" (in breve "MPEG-7"). Lo scopo è quello di definire un insieme standard di descrittori delle caratteristiche di un dato oggetto multimediale come possono essere il colore medio di un'immagine oppure l'energia media di un segmento audio. MPEG-7 inoltre vuole standardizzare delle strutture (Description Schemes o DS) di descrittori nonché un linguaggio per specificare gli schemi di descrizione (Description Definition Language o DDL). Lo standard sarà applicabile al materiale AV come immagini, grafici, modelli 3D, audio, parlato e video. Una descrizione MPEG-7 potrà far parte del documento stesso o essere disponibile in qualsiasi altro posto della terra. Quando il contenuto e la sua descrizione non sono nello stesso posto allora si rende necessario un meccanismo di link bilaterale che colleghi univocamente i due.

In Figura 1 viene mostrato un diagramma a blocchi che mostra il ruolo di MPEG-7. Questo diagramma include l'estrazione delle caratteristiche (analisi), la descrizione, ed il motore di ricerca (applicazione). L'estrazione automatica delle caratteristiche (o 'descrittori') sarà molto importante per dare una descrizione MPEG-7 completa. Tuttavia l'estrazione automatica

delle caratteristiche non è sempre possibile e le difficoltà aumentano al crescere del livello di astrazione che rappresentano.

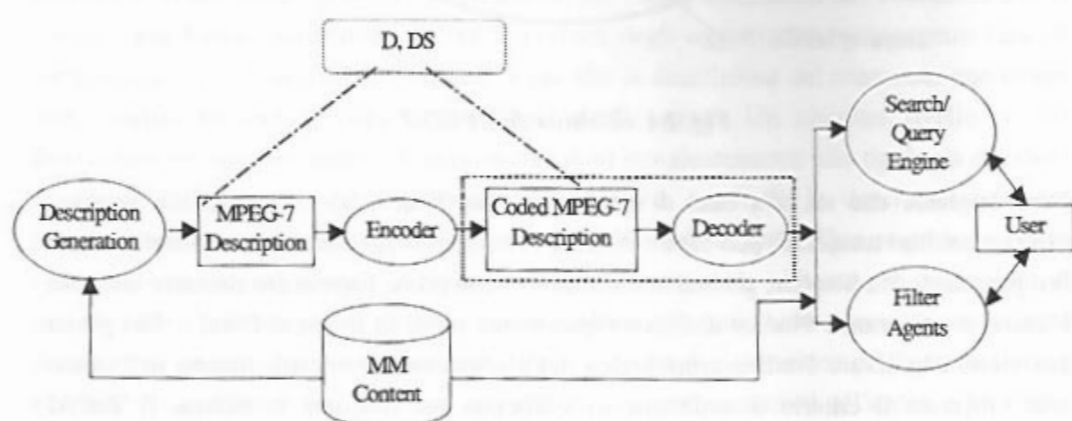


Fig. 2 Le possibili applicazioni di MPEG-7.

Nonostante la loro utilità i metodi automatici o semi automatici di estrazione delle caratteristiche esulano dagli obiettivi di MPEG7, infatti la non standardizzazione di questi ultimi non compromette l'interoperabilità e apre un nuovo mercato per le industrie del settore. Un'altra ragione per non standardizzare gli strumenti di analisi è dovuta alla elevata velocità con cui questi ultimi si evolvono diventando sempre più potenti ed affidabili. Per questi motivi anche i motori di ricerca e le interfacce utente non sono oggetto di definizione. In Figura 2 è schematizzato il piano di lavoro MPEG7. L'attenzione è posta sullo sviluppo di soluzioni nuove per la descrizione del contenuto audio-video. Spiegazioni più dettagliate su questo progetto si possono ritrovare in [6],[7].

3. ToCAI

Questo DS è composto dall'aggregazione di quattro schemi di descrizione principali: il Table of Contents (ToC), l'Analytical Index (AI), il Context ed il Meta-Descriptors.

ToC DS

Il ToC DS è composto da diversi livelli gerarchici dove al livello più basso è presente una descrizione dettagliata della struttura sequenziale del documento audio-video, mentre quello più alto da una descrizione compatta della semantica associata. Un aspetto chiave è che gli elementi di ciascun livello sono disposti in ordine cronologico.

Per esempio, per un telegiornale si può avere una ToC del seguente tipo:

- ❖ Summary (0:2'30'')
- ❖ Internal affairs (2'31':7')
 - Speaker presentation (...)
 - First reportage (...)
 - Shot 1 (...)
 - Shot 2 (...)

- ...
- Speaker presentation (...)
-
- ❖ International affairs (...)
-

dove i dati tra parentesi specificano la posizione temporale del segmento, mentre le etichette identificano la semantica. Ogni elemento della ToC in diversi modi come ad esempio un k-frame o un segmento audio. Il ToC DS è molto utile per scorrere velocemente un documento dato che fornisce un riassunto dello stesso ad ogni livello di dettaglio. Riassumendo i concetti chiave del ToC DS sono:

- gerarchia,
- ordine cronologico.

Analytical Index DS

L'AI permette di avere un insieme ordinato di oggetti audio-video. Un elemento dell'AI può puntare a differenti livelli relativi alla gerarchia definita nella ToC o in base ad altri criteri. Ogni elemento dell'AI può puntare a più oggetti audiovisivi. Questa funzionalità permette di muoversi attraverso documenti multimediali in modo non sequenziale, guardando ad esempio tutte le scene che contengono un dato oggetto. Gli elementi dell'AI possono essere delle entità semantiche, come le scene che appartengono alla categoria dialoghi, categorie di immagini come sfondi oggetti, ma anche oggetti audio come motivi musicali o parole chiave estratte dalla trascrizione del parlato. Tutti questi oggetti possono essere ordinati secondo diversi criteri, elencati nel DS. Riassumendo i concetti chiave per l'AI sono:

- ordinamento,
- puntatori di riferimento.

Context DS

Il ToCAI, che fa riferimento alla struttura di un documento AV, deve essere considerato unitamente ad uno schema di descrizione che dia informazioni sulla categoria del documento stesso. Questo descrittore del contesto include descrittori quali il titolo del programma, attori, regista, lingua, paese d'origine, etc. Queste informazioni sono molto utili per restringere il campo di ricerca.

Meta-descriptors DS

Questo DS ha la funzione di incorporare all'interno del ToCAI un insieme di descrittori che diano informazioni su come e con che grado di affidabilità un dato descrittore è stato istanziato. In altre parole non servono per descrivere il contenuto ma danno un'indicazione sull'affidabilità del valore assegnato ad un descrittore.

4. Struttura dettagliata dello schema di descrizione proposto

Si passa ora a descrivere la struttura del ToCAI presentando l'organizzazione gerarchica dei sotto schemi di descrizione e descrittori associati (vedi Figura 3). In questa fase è stata utilizzata la notazione UML (Universal Modelling Language) [4].

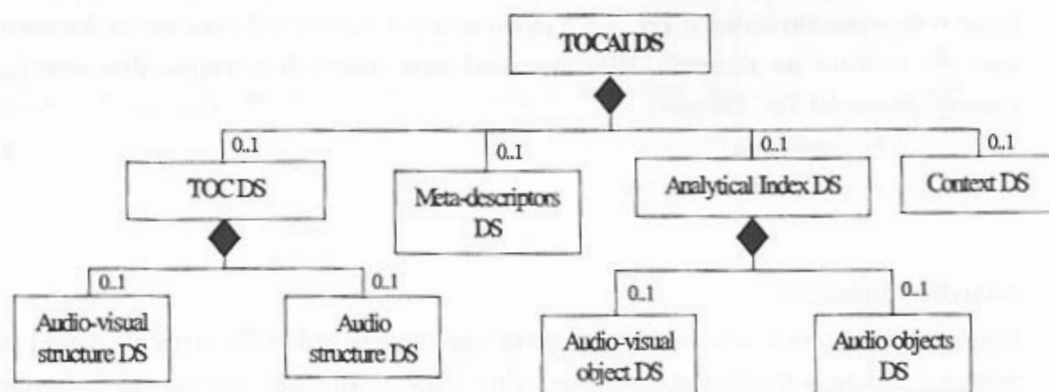


Fig. 3 La struttura del ToCAI DS.

ToC DS

Descrive la struttura temporale del documento AV a diversi livelli di astrazione. E' composto da due DS uno chiamato Audio-visual Structure e l'altro Audio Structure. L'utilizzo di un DS audio-video in luogo di un video è dovuto al fatto che, dal punto di vista semantico, è spesso necessario considerare l'informazione video assieme a quella audio associata. In tal modo si possono recuperare livelli semantici intermedi per la descrizione.

Audio-visual structure DS

Questo DS schematizzato in Figure 4. I due Time-code Ds specificano l'inizio e la fine del documento. Il cuore di questo DS è il Scene DS. Una scena è data dalla segmentazione temporale caratterizzata da una semantica coerente ad un certo livello di astrazione. Il Scene DS è composto da un certo numero di sotto scene, un riferimento temporale (due time-code DS) ed un descrittore del tipo di scena (una stringa ed un'icona). I componenti elementari di una scena sono gli shot. Questi sono rappresentati tramite il Shot DS che contiene il tipo di effetto di editing utilizzato (cut, dissolve, fade in/out, etc.) a la posizione temporale. In ultimo sono compresi un insieme di schemi di descrizione per i k-frame, mosaici relativi agli shot.

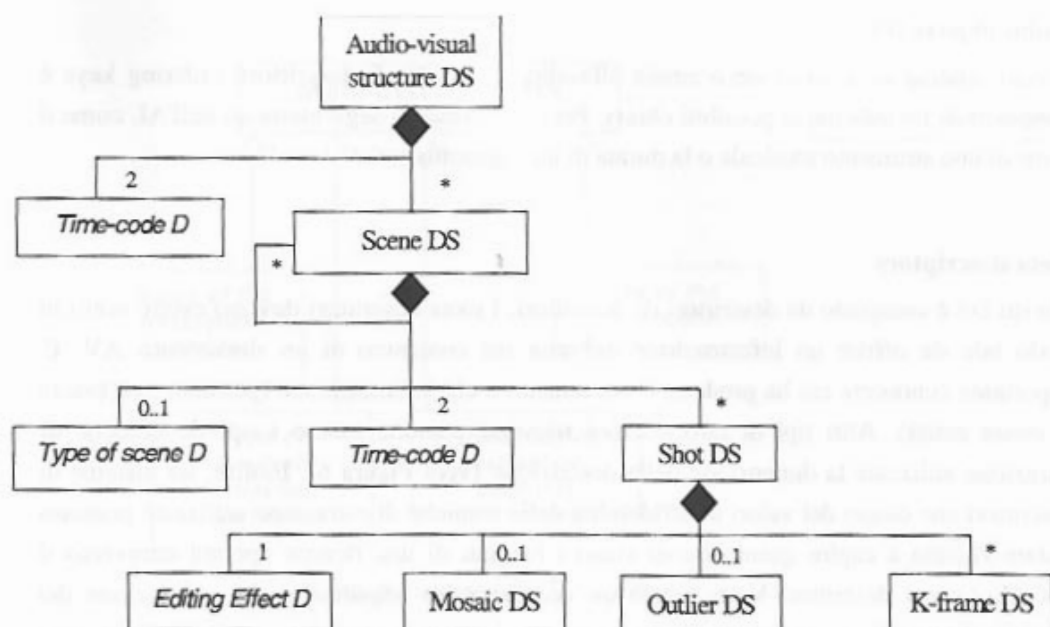


Fig. 4 L'Audio-visual structure DS.

Audio-structure DS

Questo DS ricalca la composizione del precedente. Si possono quindi avere vari livelli di scena audio. I descrittori connessi allo schema Homogeneous audio DS identificano i livelli dell'albero, dove sono presenti segmenti audio ottenuti da una sorgente omogenea (ad esempio uno stesso parlatore, un rumore, etc.). Ognuno di questi descrittori è formato da un'etichetta e da un riferimento temporale.

Analytical Index DS

Come è stato detto l'AI permette di creare un set ordinato di oggetti audio-video puntando a diverse posizioni temporali e a diversi livelli della struttura del documento. Perciò questo DS ha il ruolo fondamentale di supportare il recupero di informazioni all'interno del documento. E' composto da due ulteriori DS: l'Audio-Visual object DS e l'Audio object DS.

Audio-visual objects

La struttura di questo schema è mostrata in Figura 5. Il descrittore ordering key è l'insieme dei possibili criteri di ordinamento applicabili agli elementi dell'AI (colore, volume del suono, etc.). Sono previste due classi di oggetti AV: scene e immagini e di conseguenza due DS. Per ciascun oggetto dell'indice analitico, ci sono diversi tipi di puntatori (un oggetto può puntare a differenti livelli di astrazione) e per ogni tipo di puntatore, ci sono diversi puntatori di riferimento.

Audio-objects DS

Questo schema di descrizione è simile all'audio-visual DS. Il descrittore ordering keys è composto da un insieme di possibili chiavi. Per l'ordinamento degli elementi dell'AI, come il nome di uno strumento musicale o la durata di un segmento audio.

Meta-descriptors

Questo DS è composto da descrittori di descrittori. I meta descrittori devono essere scelti in modo tale da offrire un'informazione indiretta sul contenuto di un documento AV. E' importante conoscere chi ha prodotto il documento e chi la descrizione (possono non essere la stessa entità). Altri tipi di informazioni rilevanti possono essere i tipi di tecniche di estrazione utilizzate la dimensione della descrizione (vedi Figura 6). Inoltre, un insieme di descrittori che danno dei valori di affidabilità delle tecniche di estrazione utilizzate possono aiutare l'utente a capire quanto buoni siano i risultati di una ricerca operata attraverso il ToCAI. Questi descrittori sono perciò un complemento importante alla descrizione del contenuto dato che una ricerca effettuata su descrittori ottenuti con metodi di descrizione scadenti difficilmente potrà essere utile per capire il contenuto di un documento.

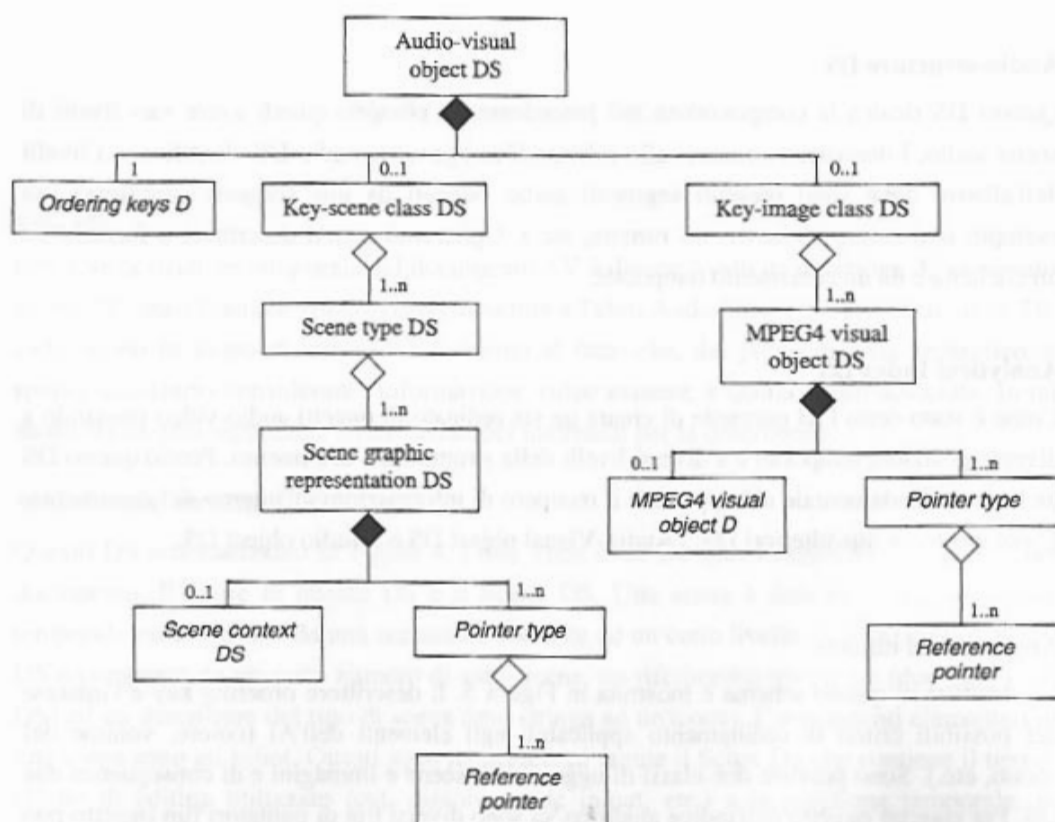


Fig. 5 L'Audio-visual object DS.

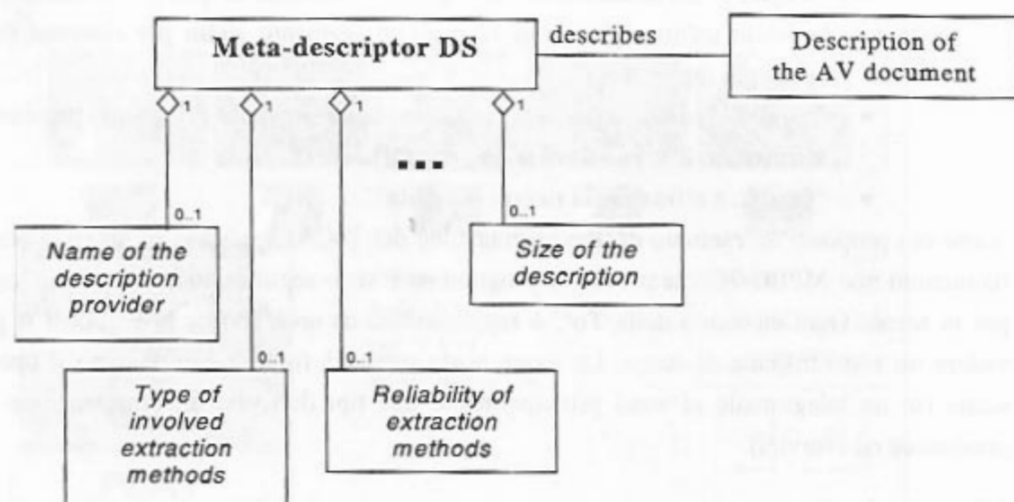


Fig. 6 Struttura del Meta-descriptors DS.

Context DS

Questo DS è composto da descrittori tipici di un programma disponibili in una guida radio televisiva (vedi Figura 7), come il titolo del programma, il paese di origine, l'anno di produzione, etc. Per esempio dalla conoscenza del regista di un film è possibile operare specifiche ricerche su questi campi.

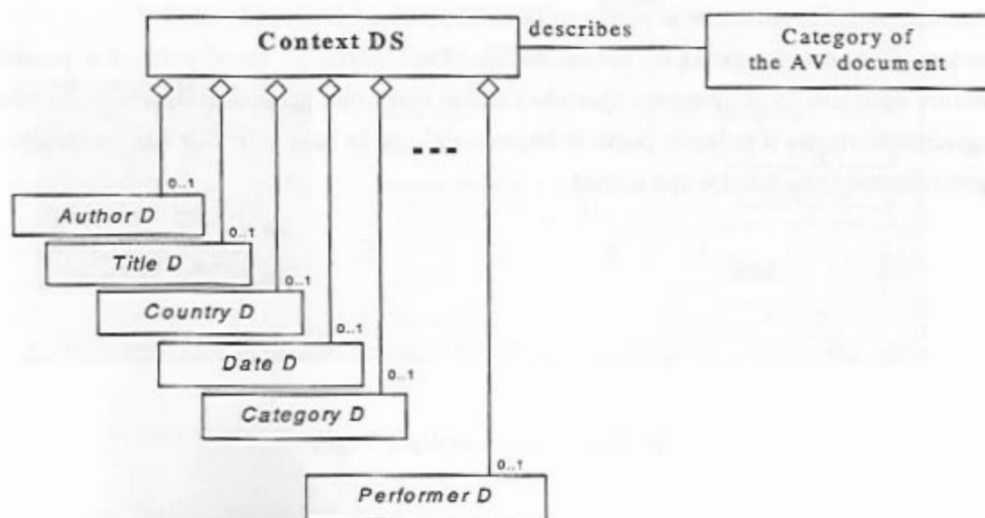


Fig. 7 Struttura del Context DS

5. Esempio applicativo

Il ToCAI ben si presta per descrivere il contenuto di lunghi programmi AV quali ad esempio i film. Il ToC permette di esplorare a diversi livelli di dettaglio (scene o shot), mentre l'AI dà la possibilità di recuperare i personaggi o sfondi particolari presenti nel filmato. Il DS proposto permette di eseguire le seguenti ricerche:

- "Voglio avere un'idea (e/o i dettagli) del contenuto di questo documento AV vedendo un'immagine o ascoltando un segmento audio per ciascuna delle scene più rappresentative".
- "Voglio vedere una lista ordinata degli oggetti principali presenti e selezionare le scene dove sono presenti".
- "Quanto è affidabile la ricerca eseguita?"

Viene ora proposto un esempio di implementazione del ToCAI applicato ad un telegiornale (contenuto neo MPEG-7Content Set). Il programma è stato segmentato in shot raggruppati poi in scene. Ogni elemento della ToC è rappresentato da un k-frame. In Figura 8 si può vedere un sotto insieme di scene. Le icone poste sotto i k-frame rappresentano il tipo di scena (in un telegiornale ci sono principalmente due tipi di scene: la presentazione del conduttore ed i servizi).

L'interfaccia utente

Lo scopo di questa interfaccia è quello di permettere l'esplorazione di un documento AV attraverso descrizioni, organizzate secondo lo schema del ToCAI, in modo semplice e naturale. Su questa base è stata sviluppata una semplice applicazione Visual Basic. Per non generare confusione nell'utente permettendogli di concentrarsi sulla parte concettuale, il numero di oggetti attivi è stato tenuto basso. Come già detto il ToCAI consiste di due parti complementari Toc e AI (Figura 9), che presentano forti analogie. Rappresentano entrambe un tipo di ordinamento, nel primo caso di tipo temporale mentre nel secondo il criterio è scelto tra quello disponibili (valore della tinta media di un immagine, etc.). Per evidenziare questa caratteristica entrambe le parti sono state rappresentate in modo simile.

Ciascuna di esse è composta da alcuni livelli ed attraverso gli stessi pulsanti è possibile esplorare ogni livello selezionato. Quando l'utente trova una parte interessante può vederla integralmentetramite il pulsante posto in basso a sinistra. In base al livello viene mostrato un oggetto diverso (uno schot o una scena).

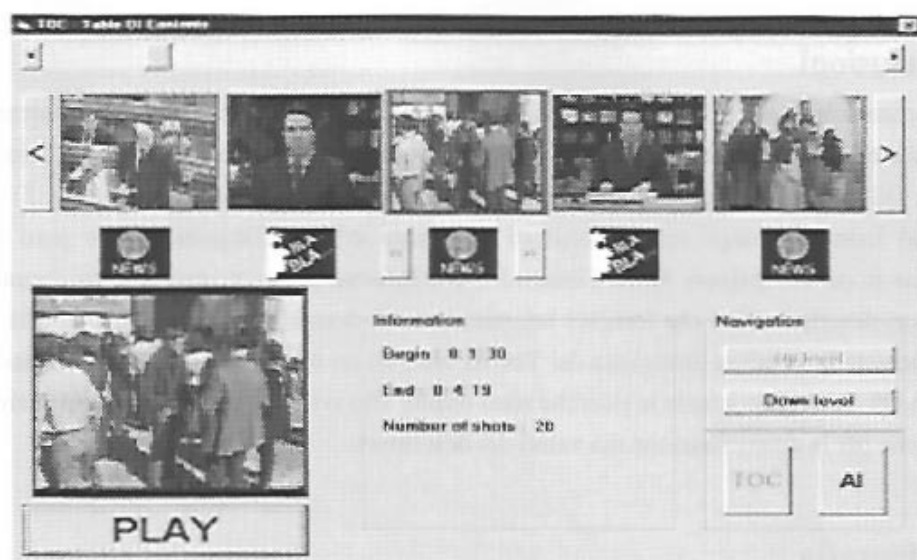


Fig. 8 Implementazione del ToC.



Fig. 9 Implementazione dell' AI.

6. Conclusioni

Questo articolo presenta il ToCAI, uno schema di descrizione del contenuto di un documento multimediale. Il DS proposto e' basato su quattro strutture principali. 1) Una Table of Contents DS che caratterizza semanticamente la struttura temporale del documento; 2) Un Analytical Index DS degli insiemi ordinati di oggetti ed i collegamenti alle parti del documento in cui compaiono; 3) Un Context DS che Informa sulla categoria del programma; 4) Un meta-descriptors DS che fornisce informazioni sui descrittori e la loro affidabilità. E' stata presentata la struttura dettagliata del ToCAI oltre ad un esempio di applicazione basata su questo DS. Allo stato attuale le ricerche sono rivolte allo sviluppo di tecniche di estrazione automatiche per la determinazione dei valori dei descrittori.

7. Bibliografia

- [1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati e L. Rossi, "The TOCAI DS for audio-visual documents. Structure and concepts", *ISO/IEC JTC1/SC29/WG11/ M4586, MPEG99*, Seoul, Korea, Mar. 1999.
- [2] N. Adami e R. Leonardi, "Identification of editing effects in image sequences by statistical modelling", *Proc. Picture Coding Symposium '99*, Portland, OR, U.S.A., Apr. 1999.
- [3] A. Ferman, A. Tekalp e R. Mehrotra, "Effective content representation for video", *Proc. IEEE International Conference Image Processing*, Chicago, IL, Oct. 1998.
- [4] M. Fowler, "UML Distilled. Addison Wesley", Longman, 1997.
- [5] O. N. Gerek e Y. Altunbasak, "Key Frame Selection from MPEG Video Data", *Proc. SPIE Visual Communications and Image Processing*, 3024:920-925, 1997.
- [6] MPEG Requirement Group, "MPEG-7: Context and objective", *ISO/IEC JTC1/SC29/WG11 N2460, MPEG98*, Atlantic City, USA, Oct. 1998.
- [7] MPEG Requirement Group, "MPEG-7: Requirements", *ISO/IEC JTC1/SC29/WG11 N2461, MPEG98*, Atlantic City, USA, Oct. 1998.
- [8] Y. Rui, T. Huang e S. Mehrotra, "Browsing and retrieving video content in a unified framework", *Proc. IEEE Workshop on Multimedia Signal Processing*, Dec. 1998.
- [9] C. Saraceno e R. Leonardi, "Indexing audio-visual databases through a joint audio and video processing", *International Journal of Imaging Systems and Technology*, 9(5):320-331, Oct. 1998.
- [10] C. Saraceno e R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing", *Proc. International Conference on Image Processing 1998*, Chicago, IL, U.S.A., Oct. 1998.
- [11] S. Smoliar e L. Wilcox, "Indexing the content of multimedia documents", *Proc. Second International Conference on Visual Information Systems 1997*, San Diego, CA, 1997.