L'innovazione possibile nella prospettiva del *MultiMedia Information Retrieval* (MMIR)

di Roberto Raieli e Perla Innocenti

1 Premessa

Quando si parla di *MultiMedia Information Retrieval* (MMIR), non si annuncia ormai alcunché di nuovo agli esperti di tecnologia dell'informazione che lavorano in ambiti molto vicini alla documentazione e al *knowledge management*, quali l'ingegneria informatica, l'intelligenza artificiale, o la *computer vision*. Nel contesto internazionale della ricerca sull'*information technology*, infatti, la sigla MMIR è ben nota a ingegneri e matematici, e indica il sistema *organico* delle tecnologie del *visual retrieval* (VR), dell'*audio retrieval* (AR), del *video retrieval* (VDR) e del *text retrieval* (TR), separatesi e sviluppatesi dalle tecniche di *information retrieval classiche*, seguendo una linea di aggiornamento e ottimizzazione delle metodologie di trattamento e recupero dell'informazione digitale.

La riflessione sullo sviluppo concettuale dell'argomento e l'interesse alla possibile rivoluzione di prospettiva metodologica e operativa, devono invece essere ancora introdotte tra coloro che operano più direttamente nella gestione dell'informazione nelle biblioteche, nelle mediateche, nei centri di documentazione o nelle biblioteche digitali. Solo pochi studiosi sono da tempo sensibili alla necessità di introdurre nell'ambito tradizionale della documentazione e della biblioteconomia la riflessione e la sperimentazione di queste nuove tecnologie[1]. Il contesto internazionale delle scienze biblioteconomiche, e l'ambito degli studi italiani tra i primi,

ROBERTO RAIELI, Biblioteca di area delle arti, Sezione spettacolo, Università Roma 3, via S. Agata dei Goti 4, 00184 Roma, e-mail raieli@uniroma3.it.
PERLA INNOCENTI, Sistema bibliotecario di ateneo-Centro sistema IT, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, e-mail perla.innocenti@polimi.it.
I capitoli 1 e 6 sono stati scritti da entrambi gli autori, i capitoli 2 e 3 da Raieli e i capitoli 4 e 5 da Innocenti.
Il presente scritto riprende i temi trattati nel recente volume *MultiMedia information retrieval: metodologie ed esperienze internazionali di content-based retrieval per l' informazione e la documentazione*, a cura di Roberto Raieli e Perla Innocenti, Roma: AIDA, 2004, cui si rimanda per i contibuti di esperti e studiosi del settore e per la bibliografia introduttiva generale (l'indice del volume è disponibile a <http://www.aidaweb.it/pubblicazioni.html>).
Data di ultima consultazione dei siti Web: dicembre 2004.

1 Come esempi, tra i principali, gli studi di Peter Enser in Inghilterra, di William I. Grosky negli Stati Uniti, e di Eric Paquet in Canada (cfr. Bibliografia di riferimento).

hanno ancora l'occasione di accogliere in tempo la discussione, al momento in cui i database e le interfacce di *MultiMedia Information Retrieval* sono in fase di sperimentazione, con la conseguente possibilità di indirizzare lo sviluppo di tali sistemi secondo le proprie necessità, e non doversi poi affaticare per ricondurre successivamente alle proprie prospettive logiche e tecniche differenti.

Già in alcune università e centri di studio del nostro paese ricercatori e tecnici stanno mettendo a punto e sperimentando sistemi informatici di MMIR, è quindi necessario anche per i bibliotecari e i documentalisti percepire quanto l'intero complesso del *MultiMedia Information Retrieval* possa costituire la base di una nuova strategia di ricerca dell'informazione, la quale, ponendosi più in avanti delle metodologie tradizionali di *information retrieval* (IR) basate sulla terminologia descrittiva, punta al reperimento basato sull'effettivo e oggettivo contenuto del documento, riprogettando le architetture necessarie nei nuovi grandi database multimediali.

La necessità e l'efficacia di un complesso organico del *MultiMedia Information Retrieval* non escludono, però, che ognuno dei suoi sottosistemi abbia un contesto specifico, sia tecnico sia metodologico. Così nell'ambito dell'*information retrieval* si prospetta un basilare rinnovamento di principio che si rende concreto in più rivoluzioni di metodo, ognuna in sviluppo con tempi e forme proprie: VR, VDR, AR, TR. Tutto questo non esclude né sorpassa, ma ricomprende a un livello più alto, la metodologia *terminologica* che è tutt'oggi il nucleo dell'IR, riservandogli ambiti, forme e modalità incontestabilmente proprie, ma includendolo organicamente nel MMIR e rinominandolo *text retrieval*. Inteso sotto la nuova prospettiva il termine stesso *information retrieval* può arricchirsi di significato, diventare praticamente sinonimo di *MultiMedia Information Retrieval*, e rendere non necessaria l'aggiunta della specifica *MultiMedia* dinanzi a *information retrieval*, potrebbe essere sufficiente parlare di *information retrieval rinnovato*.

I sistemi informativi di diverse organizzazioni pubbliche e private stanno attraversando, del resto, una fase di profonda trasformazione, dovuta alla crescente richiesta da parte del mondo accademico, scientifico e professionale di documentazione aggiornata connotata da supporti eterogenei, forme espressive multimediali e linguaggi diversi. Un contesto molto rinnovato e differente dalle tradizionali collezioni monotipologiche, che richiede specifiche modalità di gestione, accesso e fruizione dei documenti. I sistemi di MMIR promettono qui significative applicazioni, e sono già in uso in diversi settori: dall'ingegneria all'astronomia, ai GIS e ai sistemi di *remote sensing* fino ai sistemi di identificazione e di sorveglianza; dagli archivi di film e di video al giornalismo e allo spettacolo; dalla biochimica al settore medico; dal design all'architettura e all'archeologia; dai beni culturali alla didattica e all'*e-learning*; dalla documentazione aziendale all'*e-commerce* e ai dispositivi domestici.

2 Il metasistema *MultiMedia Information Retrieval*

Documentalisti e bibliotecari hanno da sempre adottato un'interpretazione *user-centered* nell'impostazione delle metodologie di *information retrieval*, focalizzando l'attenzione sui modi concettuali, interpretativi e testuali con cui gli utenti descrivono e trattano l'informazione, lasciando in secondo piano, diversamente da altri specialisti dell'informazione, le modalità automatiche di strutturazione, immagazzinamento e recupero dei dati. Nell'ultimo decennio, però, la centralità dei documenti multimediali e i nuovi mezzi offerti dalle tecnologie digitali hanno favorito la creazione di basi dati multimediali ben altrimenti complesse rispetto alle basi dati tradizionali, mettendo in discussione la piena efficacia dei principi di IR nel trattamento delle nuove tipologie di documenti.

Il crescente impiego dell'*information retrieval* in ambito tecnico e commerciale, inoltre, ha stimolato l'interesse della *computer science*, che a differenza della biblioteconomia e della documentazione ha affrontato queste tematiche impostando la propria prospettiva sulle strutture e sugli algoritmi di elaborazione dei dati informativi. Da un'ottica *computer-centered*, senza opporsi alla prospettiva *user-centered* e considerando tutte le questioni di interesse dello *user*, il problema consiste nella costruzione di indici efficaci, nell'elaborazione di *query* con un alto livello di performance e nello sviluppo di algoritmi di *ranking* che migliorino la qualità assoluta della risposta. Si evidenzia così l'inefficacia della tecnica tipica basata su un complesso di *surrogati descrittivi* dei documenti multimediali, in base alla quale, sia il sistema sia l'utente, sono forzati a operare su di essi, e si rende necessario comprendere quale sia concretamente il tipo di informazione estraibile da immagini, filmati o registrazioni sonore, nonché come tali informazioni possano essere rappresentate e organizzate per supportare richieste chiaramente orientate ai *contenuti concreti*.

Il concetto di *query multimediale*, allora, si definisce come qualcosa di differente e diversamente evoluto rispetto alla classica *query* terminologica, da impostare con riguardo alle caratteristiche oggettive e specifiche dei materiali immagazzinati nel sistema interrogato, il quale deve essere implementato con sistemi di gestione e accesso integrati per documenti tipologicamente eterogenei, attraverso l'impiego di specifici sistemi per l'indicizzazione, la ricerca e l'estrazione automatica di dati rappresentativi del complesso contenuto dei documenti multimediali[2]. Inoltre, poiché la qualità del recupero delle informazioni è largamente influenzata dall'interazione dell'utente con il sistema, anche nei confronti dell'utente molto deve cambiare, e il metodo di approccio ai *database* multimediali deve essere riformulato sulla base delle più complesse esigenze di definire la *query* con dati *visivi* e *sonori* e non soltanto con dati terminologici. Alle classiche interfacce dei database testuali, dunque, le quali consentono la ricerca in un indice composto esclusivamente di termini estratti dai documenti o inseriti in metadati testuali, devono succedere interfacce che permettano di formulare la *query* in diverse dimensioni, non solo tramite i termini ma anche attraverso le immagini e i suoni, rendendo in tal modo possibile la ricerca in indici composti da testi estratti dalle didascalie o dal parlato, da immagini chiave di una sequenza, da semplici figure, da melodie, da forme, colori e suoni, senza con ciò escludere l'importanza che continuano a mantenere i dati testuali, descrittivi o classificatori, relativi ad aspetti non specificamente audiovisivi del documento.

Seguendo questa tendenza, al posto dei tradizionali sistemi di indicizzazione e ricerca detti *term-based*, basati sull'impiego di termini descrittori, sono via via stati sperimentati progrediti sistemi di archiviazione e recupero definiti *content-based*, in cui i *descrittori* sono dei veri e propri *metadati*, caso per caso della stessa natura dei dati cui si riferiscono, e di cui possono consentire nel modo più funzionale l'analisi e la ricerca. Il metodo *content-based* risulta quindi pienamente efficace per cogliere l'obiettivo del *MultiMedia Information Retrieval*: restituire l'oggetto che esattamente si cerca, al di là di ogni vincolante mediazione classificatoria, trattando, immagazzinando e richiamando ogni genere di documento digitale tramite gli elementi di linguaggio, o di *metalinguaggio*, propri dello specifico contenuto.

Attualmente è necessario distinguere tra i sistemi di *information retrieval term-based*, basati su informazioni testuali per la ricerca di documenti testuali o indifferente-

2 William I. Grosky, *Managing multimedia information in database systems*, «Communications of the ACM», 40 (1997), n. 12, p. 73-80.

mente di documenti multimediali, e i sistemi di *information retrieval content-based*, basati su dati visivi per la ricerca di documenti visivi, dati audiovisivi per la ricerca di audiovisivi e dati sonori per documenti sonori.

In prospettiva invece, in un unico ambito di IR *rinnovato* coincidente con il MMIR, si potranno, auspicabilmente, considerare solo sistemi di *information retrieval content-based* o più propriamente di *MultiMedia Information Retrieval*, dove l'attributo *content-based* ricomprende naturalmente al proprio interno anche *term-based*, dacchè se l'obbiettivo è il *contenuto* non importa a priori se questo sia terminologico, visivo eccetera. Tali sistemi si specializzeranno in:
– sistemi di *text retrieval*, nei quali il principio stesso del *content-based information retrieval* legittima e garantisce l'utilizzo di dati terminologici e testuali per la ricerca di informazione testuale;
– sistemi di *visual retrieval*, in cui i file di immagini fisse sono cercati e recuperati tramite dati visivi interni al file, quali ad esempio colore, contorno, struttura, forma, orientamento e distribuzione spaziale;
– sistemi di *video retrieval*, dove per il recupero di documenti audiovisivi si utilizza il linguaggio audiovisivo, cioè elementi di ricerca ricavati dalle immagini del filmato, dal movimento degli oggetti nelle inquadrature, dall'analisi degli stacchi di montaggio o della traccia sonora;
– sistemi di *audio retrieval*, nei quali l'informazione sonora è ricercata tramite suoni, ricavando quindi i dati di *query* dall'analisi dei tempi, delle frequenze, dei ritmi o delle melodie.

Insieme a ciò, avendo riguardo del senso del complesso omogeneo dei quattro metodi di *MultiMedia Information Retrieval*, si deve precisare che per raggiungere un buon livello di soddisfazione nel recupero dei documenti da un *database* multimediale è necessaria la compresenza di tutte le modalità, e in particolare di quella basata sui termini. L'interrogazione terminologica può, per prima cosa, essere un ottimo metodo preliminare per selezionare una parte della grande quantità di documenti di un archivio, e per centrare la ricerca in base a dati quali le tipologie, le classi, i titoli, gli autori; successivamente, essa può costituire un sistema finale di ripulitura dall'inevitabile *rumore* specifico di un'interrogazione *content-based*, precisando quello che il sistema non è in grado di avvertire nell'analisi diretta delle caratteristiche rappresentative del documento. Soprattutto, però, tutti i procedimenti possono operare in costante interazione, in un unico sistema e con un'unica schermata di ricerca, nella composizione di una formula di *query* che combinando immagini e testi, suoni e testi, o tutti insieme, possa servire per la ricerca di documenti molto complessi, il cui contenuto informativo si estende a tutti i livelli di *senso* e *significato*, dove le definizioni concettuali hanno molta importanza.

2.1 I principi dell'indicizzazione *content-based*

I dati e i documenti multimediali possono essere definiti *oggetti multimediali* di vario genere, che per la loro complessa struttura non sono efficacemente rappresentabili nei sistemi *term-based*. Tali oggetti sono ripresi dal mondo reale, estratti anche con strumenti di registrazione diretta – fotografica o sonora, magnetica o digitale – e tradotti in rappresentazioni dell'*oggetto reale*. Con le tecniche attuali – ad esempio, *image processing* o *speech recognition* – i sistemi elettronici riescono a identificare adeguatamente gli oggetti reali, estraendo determinate informazioni dal corrispondente oggetto multimediale. Tali informazioni sono dette *features* e sono contenute nel *data model* dell'oggetto multimediale, insieme di dati meno complesso di quello dell'intero file proprio dell'oggetto che rappresenta. Dagli *oggetti reali*, dunque, apparte-

nenti al mondo dell'esperienza quotidiana, vengono ripresi gli *oggetti multimediali*, documenti di vario genere di questa esperienza trattabili dai sistemi informativi, e da questi ultimi sono poi estratti i *data model*, insiemi di dati digitali identificativi del documento che possono essere trattati dai sistemi informativi più avanzati. Tale processo, che dalle cose della realtà quotidiana porta al loro trattamento documentario tramite i *data model*, è chiamato *multimedia data modelling*, e si propone come il nucleo teorico e tecnico del *MultiMedia Information Retrieval*[3].

Il *data model content-based* rappresenta le proprietà delle cose, le loro relazioni e le operazioni definite su esse, e tali *concetti astratti*, nondimeno, sono in esso tradotti in dati digitali, *fisicamente situabili* nel sistema del *database*. Così, attraverso la mediazione del *data model*, le *query* e altre operazioni sugli oggetti reali possono essere trasformate in operazioni sulle rappresentazioni astratte di tali oggetti, che sono a loro volta trasformate in operazioni sui dati digitali che traducono le rappresentazioni astratte nel *linguaggio* del sistema elettronico. Tutte le operazioni relative agli oggetti del *database* multimediale si basano su tali modelli rappresentativi: per le immagini, ad esempio, i *data model* possono contenere dati come la risoluzione, il numero di pixel, i valori dei colori, i valori caratteristici della struttura, o del tratto; durante la *query*, inoltre, si possono aggiungere ai modelli dati quali la combinazione di figure o settori di un'immagine con quelli di un'altra, lo sviluppo di una figurazione in un'altra con date caratteristiche eccetera.

Il *data modelling* realizza dunque il processo di *indicizzazione content-based* più appropriato alle caratteristiche del materiale multimediale. Il concetto di indicizzazione va inteso però, in questo contesto, in senso più largo rispetto alla sua accezione comune. Esso, nell'ambito dei documenti multimediali, va riferito a una tecnica di creazione dell'*indice* del *database* tramite l'estrazione da documenti non testuali di elementi che né sono termini né sono traducibili in termini. L'indice viene creato, quindi, impostando come collegamenti di accesso ai documenti i dati costitutivi del loro stesso contenuto multimediale: forme, colori, suoni, e altri simili elementi.

2.2 Le modalità della ricerca *content-based*

Il processo di reperimento *content-based* in un *database* multimediale è una procedura ben più complessa rispetto a quella di una banca dati testuale. Il *browsing*, video o audio che sia, assume una particolare importanza in questo ambito, le stringhe di interrogazione possono contenere direttamente valori del *data model* o oggetti multimediali campione, e infine i risultati di tali *query* non sono basati su un esatto *matching* tra i dati inviati e quelli ritrovati ma solo su certi gradi di paragonabilità e *similitudine*.

Nel procedimento più tipico l'utente avvia la ricerca con la preliminare selezione di un archivio, o di una sua parte, tramite l'uso di stringhe di testo appropriate. In questo caso, con valore propedeutico alla ricerca più propriamente multimediale, l'utilizzo di termini, di titoli e di nomi può essere il metodo più adatto e veloce per una prima riduzione della mole di tutto il materiale potenzialmente interrogabile. Si continua la ricerca con un *browsing* esplorativo, tramite cui si possono inviare al sistema le richieste semplicemente selezionando con il mouse un oggetto o una sua parte. Così, muovendosi tra tanti oggetti che assomigliano a quello cercato o che con esso sono in relazione, si possono inviare all'elaboratore diversi *data model*, contenenti i dati caratteristici da rintracciare negli oggetti dell'archivio per estrarli come risultati della *query*. Tale metodo è quello attualmente più applicato per stringere la ricerca attorno al documento che si desidera.

3 William I. Grosky, *Managing multimedia information in database systems* cit., p. 74-75.

Nei *database* più avanzati, oltre a inviare i parametri di *query* tramite la selezione di un oggetto già posseduto dal sistema, si possono compilare delle specifiche *griglie* definendo liberamente i parametri, oppure il sistema fornisce la possibilità di inserire dall'esterno dei *modelli* di esempio che verranno analizzati derivandone il *data model*. Con tale metodologia il database potrà, ad esempio, essere in grado di soddisfare richieste *content-based* come quella di trovare con l'immagine del particolare di un'opera di un pittore l'opera intera o altre opere o quelle della stessa scuola; oppure in base all'ecografia di un organo malato si potranno reperire altre immagini di organi con lesioni simili; o con le note principali di un motivo trovare diverse composizioni che lo hanno sviluppato.

Nell'insieme del *MultiMedia Information Retrieval*, definibile, per riassumere, come un *metasistema*, si differenziano aspetti tecnologici e metodologici specifici, riferibili a sistemi di *retrieval* più settoriali, dato ovviamente che nel circuito dell'informazione non circolano solo documenti pienamente multimediali, ma fanno parte di esso anche documenti specificamente visivi, audiovisivi o sonori. È necessario quindi trattare separatamente del *visual retrieval*, del *video retrieval* e dell'*audio retrieval*, come parti autosufficienti del *tutto organico* del MMIR.

Queste metodologie specifiche di *content-based retrieval* possono essere senz'altro distinte anche su base teorica, oltre che su quella tecnologica, e si sono di fatto distinte sulla base pratica della diversa attenzione ricevuta dagli studiosi e dai tecnici che le hanno sperimentate, nonché relativamente ai diversi modi e tempi di accoglimento nell'ambito applicativo e commerciale.

Così, si può parlare del *visual retrieval* come il settore *pioniere*, essendo i primi studi stati sviluppati circa vent'anni fa relativamente al trattamento delle immagini fisse. L'ambito del *video retrieval* si è aperto subito dopo e per conseguenza diretta, ma ha poi suscitato un maggiore interesse nella ricerca industriale applicandosi alla tipologia comunicativa, spesso maggiormente promossa, delle immagini in movimento, senza trascurare comunque le problematiche di base delle immagini fisse. L'ambito dell'*audio retrieval* deve invece essere presentato come il settore più giovane, e ancora in fase di studio, nonostante la grande crescita di interesse nell'ultimo quinquennio per le applicazioni pratiche del sistema.

3 *Visual retrieval*

Nella pratica tradizionale dell'*information retrieval* il trattamento dei documenti visivi è strutturato secondo le logiche *term-based*, tutto viene tradotto in termini: l'immagine, il suo contenuto oggettivo, le chiavi d'accesso che la identificano nell'archiviazione, e di conseguenza l'impostazione della *query* nella ricerca. Ma dalla rappresentazione terminologica del contenuto figurativo nasce una serie di problemi che si riflettono nella strutturazione dei sistemi di recupero[4].

Anzitutto, l'oggetto visivo viene costretto, tentando di tradurlo, nelle forme terminologiche caratteristiche della tecnica documentale in uso, e in questa prospettiva la *query* espressa testualmente punta all'incontro con i *surrogati* testuali dei documenti visivi – parole chiave, termini di indicizzazione, titoli o didascalie – a cui è meccanicamente collegata l'immagine archiviata. Oltre all'evidente inefficacia di tali forzature, l'indicizzazione di pur tutti i termini implicabili nella descrizione della vasta complessità di un'immagine non potrà mai essere esaustiva, anche con centi-

4 Peter Enser individua e discute i principali tra questi problemi e le relative conseguenze (cfr. Peter G.B. Enser, *Pictorial information retrieval. Progress in documentation*, «Journal of Documentation», 51 (1995), n. 2, p. 126-170.

naia di parole. Spesso, inoltre, le qualità di un oggetto visivo non rientrano in alcuna categoria linguistica, e risultano non solo inclassificabili ma anche terminologicamente inesprimibili. Non sempre, poi, i concetti estraibili dai documenti visivi interessano maggiormente del loro contenuto concreto, delle *rappresentazioni* in se stesse, dei tratti, delle forme, dei colori.

Considerevole è anche il problema della cultura specifica indispensabile in chi dovrà preparare o usare una struttura di base indicizzata, per cui spesso è essenziale la conoscenza di una precisa terminologia tesaurale. Il vocabolario degli studiosi risulta spesso poco intuitivo e di non facile uso per gli utenti medi e non specialisti, ma anche per gli specialisti si scopre limitativo quando si vuole andare oltre i tradizionali schemi di studio. Infine, l'estrazione in buona parte manuale dei termini descrittivi ripropone la problematica della mediazione umana, dacché nonostante l'uso di sistemi di guida e di assistenza computerizzata l'essere umano è sempre portato a decidere in base ai propri punti di vista. A ciò vanno aggiunti i problemi relativi ai costi e al tempo necessari per consentire tali delicate operazioni.

Non è difficile dunque stabilire quanto nei *database* di immagini possano risultare riduttivi e inefficaci i metodi di indicizzazione e di ricerca basati sulle annotazioni terminologiche. Negli archivi dove il contenuto dei documenti è sostanzialmente un testo, appare ovvio e appropriato che le chiavi che ne consentono l'accesso siano parole e frasi, o termini e codici, estratti *dall'interno* di quel contenuto stesso. Negli archivi di immagini invece si rivela semplificativo e impreciso attribuire, *dall'esterno*, una descrizione testuale a contenuti che si fondano su un diverso regime di senso (vedi ad esempio fig. 1), non diversamente da come verrebbe definito senz'altro improduttivo il caso opposto della ricerca tramite immagini di un contenuto testuale.

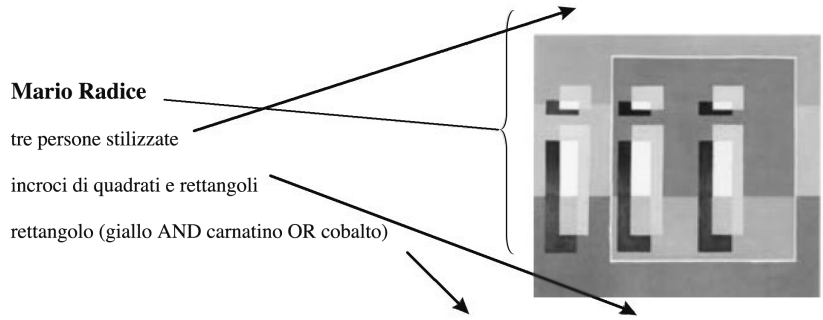

Fig. 1. Esempio di ricerca testuale-visiva

La rivoluzionarietà dei sistemi di *visual retrieval* si fonda sulla base di una tecnologia di archiviazione e recupero che tratta direttamente il contenuto visivo oggettivo dei documenti, definita per questo *content-based*, in opposizione ai tradizionali sistemi *term-based* di indicizzazione e ricerca basati su termini descrittori di tale contenuto visivo. Il metodo del *visual retrieval*, in definitiva, sperimenta e concretizza la possibilità di reperire le immagini tramite gli appropriati mezzi del linguaggio visivo stesso, come la somiglianza, l'approssimazione e i rapporti di misure e valori, utilizzando quali chiavi di ricerca figure, strutture, forme, tratti, linee e colori (vedi ad esempio fig. 2).

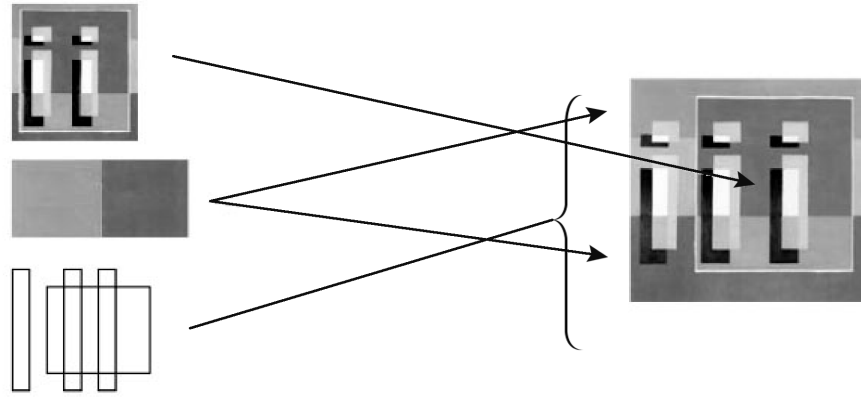

Fig. 2. Esempio di ricerca visivo-visiva

Le procedure automatiche di indicizzazione *content-based* proprie della struttura dei sistemi di *visual retrieval* evitano ogni passaggio di mediazione terminologica *vecchia maniera*, trattando in modo diretto le caratteristiche dell'oggetto originale, o meglio, i dati della sua versione documentale digitale. I dati dell'indice di ricerca, inoltre, possono essere prodotti direttamente dal sistema che li utilizzerà, quindi nella forma a esso sicuramente più idonea.

Il metodo di base del processo di trattamento automatico dei file consiste in una procedura di decomposizione strutturale e analisi dei valori che riduce la figura in un set di dati, da intendere come *modelli* o brevi abstract elettronici, ottenuti in base ad algoritmi dei valori dominanti in ogni immagine, senza interventi d'interpretazione umani. Eventuali errori e approssimazioni degli algoritmi sono dovuti a cause note, e quindi sono calcolabili come errori sistematici, che potranno essere tenuti in conto nel gestire il risultato finale. Rispetto alle variabili individuali dei metodi manuali, e agli errori interpretativi nascosti, i processi dei sistemi automatici risultano quindi spesso di maggiore affidabilità, almeno in certi contesti.

Per attuare l'indicizzazione e le successive operazioni di recupero delle immagini digitali, e per l'elaborazione dei dati e dei valori relativi a elementi e proprietà di esse, è fondamentale che il sistema, così come il sempre necessario operatore umano, siano in grado di produrre un certo livello di *astrazione* del documento. Un'astrazione che, ovviamente, non ha nulla a che vedere con l'astrarre concetti e soggetti, astrazione significa, piuttosto, *estrazione* e *modellizzazione* di alcuni elementi del contenuto oggettivo dell'immagine. Le operazioni, infatti, potranno meglio attuarsi trattando tale forma *astratta*, un *data model* distanziato dal documento immediatamente e indistintamente preso.

Si possono in genere attuare cinque differenti livelli, o modalità, per indicizzare, archiviare, ricercare e recuperare i documenti visivi digitali. Tali modalità possono essere distinte in semantica, formale, strutturale, coloristica e parametrica, e possono costituire la struttura del sistema singolarmente o in diverse combinazioni tra loro:
– l'astrazione *semantica* consiste nella definizione di un unico e preciso contenuto semantico dell'immagine, e non è una modalità tra le più tecnologicamente avanzate. Dal punto di vista dei sistemi avanzati, però, la determinazione e l'assegnazione di un'etichetta semantica è un processo ben diverso da quello tradizionale, può

infatti essere il sistema stesso a derivare tali etichette dall'analisi e dal riconoscimento delle forme e delle strutture delle figure.
– la modalità *formale* si basa sulla capacità dell'elaboratore di attuare un confronto tra le forme estratte dalla figura archiviata e quelle estratte dal modello con cui si definisce la *query*, messo a disposizione dal sistema oppure immesso dall'esterno. Il recupero per similitudine avverrà tramite il confronto dei parametri relativi a una data presenza di linee rette o curve, o loro combinazioni, in uno spazio 2D o 3D.
– il modo di astrazione *strutturale* si basa sulla scomposizione delle immagini archiviate in sezioni, il sistema stimerà poi la somiglianza della composizione strutturale di queste con la struttura delle sezioni di una figura modello, le quali faranno dunque da chiavi di ricerca. Il recupero deve avvenire secondo un certo grado di vicinanza dei dati della figura campione, o di una griglia di *query* compilata con i valori che interessano, con quelli del set di dati collegato a un documento.
– secondo la modalità *coloristica* l'immagine può essere considerata dal punto di vista dei colori, o dei diversi grigi, che la compongono, e indicizzata con i valori del proprio *range dinamico*, che è parte del file di dati. Le operazioni di archiviazione e recupero si baseranno in conseguenza sul trattamento e il confronto dei valori dei dati relativi alle proprietà coloristiche della figura.
– la modalità *parametrica* è fondata sulla determinazione dei valori dei parametri rappresentativi della forma, della struttura e del colore dell'immagine. Il sistema potrà recuperare un'immagine tramite il confronto tra i valori dei vari parametri immessi nella *query*, attraverso una figura modello o compilando un'apposita griglia, e quelli posseduti dalle immagini in archivio.

3.1 Funzionamento del sistema

Qualunque sia la modalità di trattamento dei documenti, e quindi di strutturazione del *database*, in genere il procedimento più frequente per la creazione di un sistema di *visual retrieval* è quello di analizzare e indicizzare le immagini soltanto all'atto della costituzione del *database*, o dell'aggiornamento dell'archivio.

I passaggi per la creazione dell'archivio e dell'indice si possono sintetizzare nei seguenti:
– anzitutto l'*analisi visiva*, necessaria per l'individuazione degli elementi figurativi del documento, la quale può svolgersi automaticamente o manualmente *computer-assisted*.
– quindi l'*immagazzinamento*, cioè la creazione e registrazione del file di dati generale dell'immagine, o *datafile*.
– segue la *caratterizzazione*, estrazione dei dati caratteristici relativi agli aspetti principali della figura, con conseguente creazione del *datamodel* e collegamento al *datafile* generale.
– l'*indicizzazione visiva* consiste nel successivo aggiornamento dell'*inverted* file, costituito dai dati visivi caratteristici e da quelli generali di ogni immagine.
– ultimo passaggio, ma non meno importante, è la *descrizione*, cioè la classica estrazione delle opportune informazioni testuali e il collegamento al *datafile* tramite gli schemi dei metadati.

In un sistema così creato la ricerca viene solitamente impostata a partire da una prima consultazione del *database*, quasi sempre di tipo semantico, che consenta di estrarre da esso immagini che possano poi diventare dei modelli, utilizzabili per lanciare la *query* in altre forme, avvalendosi in sostanza di una sorta di *tesauro visivo* interno all'archivio. Le immagini estratte volta per volta possono essere modificate nelle

caratteristiche, prese per parti, o associate tra loro, secondo gli strumenti che il sistema offre, rilanciando così, ogni volta, diversi modelli che centrino meglio la ricerca.

In sintesi le fasi della ricerca e del recupero consistono nei seguenti punti:
– una *ricerca preliminare*, che consiste in un'interrogazione di tipo terminologico per selezionare una parte dei documenti dell'intero *database*, tramite scorrimento di indici o inserimento di termini.
– la *ricerca visiva* propriamente detta, tramite l'utilizzo di qualcuna delle immagini raggiunte, e selezionate tramite *browsing,* come modello di esempio per lanciare la *query* visiva.
– il *matching*, cioè la cattura automatica dei documenti la cui similitudine con il campione è di grado compreso nel parametro di recupero impostato.
– il finale *approfondimento*, utilizzando ulteriori immagini estratte, modificandone le caratteristiche, selezionando parti di esse, associando le figure, per rilanciare la *query* in vario modo.

Nei sistemi di *visual retrieval* più avanzati le immagini possono essere analizzate anche in fase di *query*, quindi i modelli per l'interrogazione si possono immettere dall'esterno, o disegnare con gli strumenti a disposizione, non solo in forma di parametri ma come compiute figure di esempio. Le ricerche in questo caso possono essere condotte molto liberamente, senza i vincoli di un tesauro visivo precostituito consistente nello stesso archivio. È necessario prevedere però che la tecnologia del sistema sia in grado, in ogni fase, di analizzare e di elaborare automaticamente e in breve tempo le immagini esterne proposte.

Schematizzando il funzionamento di un sistema di *visual retrieval* avanzato si notano:
– la possibilità dell'a*nalisi visiva in fase di query*, effettuata automaticamente e non soltanto all'atto della creazione o dell'aggiornamento del *database*,
– l'uso di *modelli esterni*, che consente l'interrogazione tramite tutti i tipi di file d'immagine proponibili dall'esterno del sistema,
– gli strumenti per la *composizione di modelli*, con possibilità di produrre o disegnare liberamente un campione per la *query* tramite apposite funzioni del sistema.

La forma più pura di *visual query*, indipendentemente dagli altri mezzi di completamento, esplora direttamente la natura figurativa dell'immagine. L'interrogazione *content-based* può avvenire tanto attraverso l'impiego di modelli o campioni, quanto tramite griglie di valori da compilare appositamente, subito dopo la preselezione terminologica di una parte d'archivio utile, oppure, nei sistemi più avanzati, può essere il primo approccio all'intero *database.*

Si deve però constatare che attualmente la *visual query* interessa spesso solo in via sperimentale, i metodi automatici e *content-based* non sempre sono i più adatti a soddisfare le esigenze più elevate degli specialisti dei diversi settori di applicazione. Il senso di un documento visivo deve essere quasi sempre colto nella sua totalità, considerando simultaneamente le tante qualità visive e intellettuali, d'aspetto e di significato, concrete e astratte, e i sistemi di ricerca visivi si dimostrano inadeguati a indicare la molteplicità di punti interpretativi intellettuali utili per l'accesso alle immagini. I sistemi di *visual retrieval*, quindi, mantengono sempre validità nel caso di un approccio diretto e *visivo-contenutistico* all'immagine, ma presentano una certa limitatezza nell'approccio teorico e *intellettuale-interpretativo*. I procedimenti terminologico e visivo devono dunque operare in costante interazione, nella composizione di una formula di *query* che per la ricerca di immagini molto complesse combini opportunamente elementi sia figurativi sia testuali.

È chiaro che tutto ciò vale a maggior ragione nel campo dell'arte. Nei settori scientifici, dove le sottigliezze concettuali relative ai significati delle raffigurazioni non hanno peso dinanzi all'oggettività visiva della figura, tali problemi non si pongono, se non riguardo a questioni di interpretazione, pur sempre oggettiva, del senso reale e correttamente documentativo dell'immagine. I sistemi di *visual retrieval*, infatti, si dimostrano già molto efficaci e utilizzabili con soddisfazione in campi quali, ad esempio, l'ingegneria o la medicina, la geografia o l'astronomia.

3.2 Alcune esperienze di *visual retrieval*

Il progetto QBIC (Query By Image Content) dell'IBM, avviato alla fine degli anni Ottanta, è sicuramente il più storico dei sistemi di *visual retrieval*, ed evolvendosi di pari passo alle tecnologie dell'*image processing* rappresenta tutt'oggi uno dei sistemi più all'avanguardia. QBIC si applica ai *database* di immagini e a quelli di video, è strutturato per il trattamento di documenti prodotti in differenti campi di applicazione, ed è implementato nelle banche dati di vari tipi di istituti o aziende in molti paesi. Esso consente interrogazioni per forma, struttura, colore, tabelle parametriche, termini e combinazioni di tutte le modalità; è possibile proporre campioni dall'esterno e produrre modelli con gli strumenti messi a disposizione; inoltre, vari strumenti di elaborazione delle immagini recuperate consentono di modificarle per rilanciare la *query*; il sistema è anche in grado di riconoscere singole figure di un complesso e utilizzarle isolatamente per le interrogazioni, o combinarle con quelle di altri complessi.

Una delle più recenti e prestigiose implementazioni di QBIC è quella presso la collezione digitale del Museo dell'Hermitage, dove agli utenti è data la possibilità di effettuare le ricerche visive in base al colore, dosando le tonalità su uno spettro definito, e in base alla forma, elaborando un modello su una tabella formale-coloristica[5].

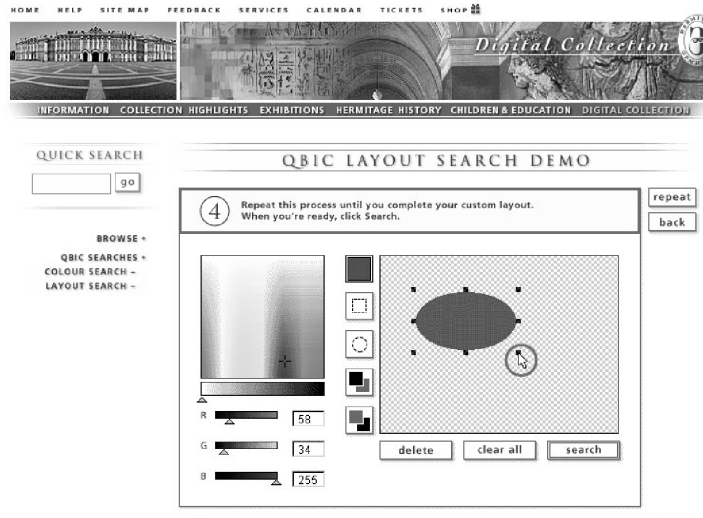

Fig. 3. Demo di QBIC sulla collezione digitale dell'Hermitage

5 Il Museo dell'Hermitage mostra in rete una demo del nuovo sistema di ricerca della collezione digitale: <www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English>. Nell'*home page* di QBIC si trovano informazioni anche su altre tecnologie e progetti IBM correlati: <wwwqbic.almaden.ibm.com>.

In continua fase di sviluppo è il sistema Zomax, realizzato dall'ISIS (Intelligent Sensory Information Systems) research group dell'Università di Amsterdam. Il sistema è stato creato per l'utilizzazione nel Web, con la documentazione liberamente disponibile in Internet, ed è composto da due parti principali: PicToVision, modulo di *image processing* che consente la segmentazione, l'equalizzazione e altre operazioni di trattamento delle immagini rintracciabili in rete; e PicToSeek, vero e proprio modulo per l'indicizzazione e la ricerca *content-based*. PicToSeek è stutturato per estrarre, indipendentemente dalla tipologia specifica dell'immagine, le caratteristiche invarianti proprie dei singoli documenti visivi del *database*, le quali sono poi confrontate con il set di caratteristiche invarianti derivato dall'imagine di *query*, che può essere proposta e caricata nel sistema semplicemente inserendo nell'interfaccia il relativo URL[6].

Fig. 4. Schermata con risultati di ricerca di PicToSeek

Altro importante programma di *visual retrieval* è il Columbia's content-based visual query project, realizzato dall'Image and ATV Lab della Columbia University di New York. Il programma è diviso in diverse sezioni, ognuna delle quali è predi-

6 Il sistema Zomax è disponibile per demo complete all'indirizzo <www.wins.uva.nl/research/isis/zomax>.

sposta per rispondere a necessità di ricerca diverse. Il modulo principale è denominato VisualSEEk, in esso è possibile impostare le *query* in base al colore e al contorno delle figure, nonché utilizzare strumenti per la creazione di modelli ed esempi. Il modulo WebSEEk è quello di impiego più semplice, basato sui testi e sui colori, realizzato per essere applicato anche nel Web. MetaSEEk, infine, passato poi nell'ambito dell'IMKA (Intelligent Multimedia Knowledge Application) project, è la sperimentazione di un meta-motore di ricerca applicabile ad altri motori di ricerca *content-based*[7].

In Italia, una rilevante applicazione della tecnologia del *visual retrieval* è quella del sistema QuickLook, messo a punto dal DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione) dell'Università di Milano-Bicocca e dall'ITC (Istituto per le tecnologie della costruzione) del CNR di Milano. Il sistema di ricerca, che si applica alle immagini fisse e ai video, è in grado di combinare interrogazioni alfanumeriche relazionali, interrogazioni basate sul contenuto visivo estratto automaticamente, e interrogazioni basate sulla similarità testuale tra descrizioni associate agli elementi visivi. È possibile interrogare l'archivio secondo diverse strategie e raffinare progressivamente la risposta del sistema indicando la rilevanza o la non rilevanza degli elementi visivi reperiti, con la possibilità di interagire introducendo vari parametri di aggiustamento della mira, i quali analizzati volta per volta da QuickLook producono ulteriori indicazioni sulla rilevanza dei documenti recuperati nell'archivio[8].

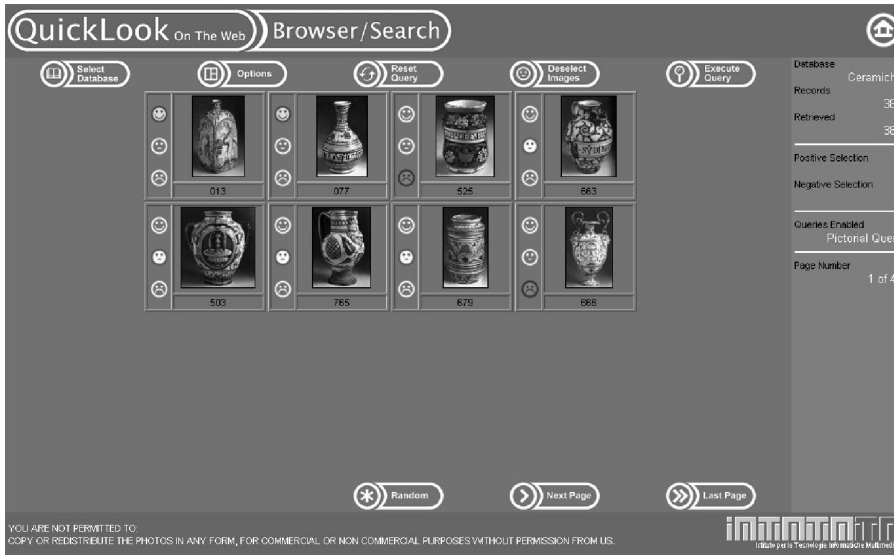

Fig. 5. Schermata della demo di QuickLook

7 Tutti i moduli del Columbia's content-based visual query project sono disponibili in rete per una prima sperimentazione: VisualSEEk all'indirizzo <www.ctr.columbia.edu/VisualSEEk>, WebSEEk all'indirizzo <www.ctr.columbia.edu/WebSEEk>, e il MetaSEEk Project al <www.ctr.columbia.edu/MetaSEEk>.

8 *QuikLook* è presentato all'URL <quicklook.itc.cnr.it>.

Tra i progetti italiani, rilevante è anche il sistema JACOB (Just A COntent Based query system for video databases), messo a punto nel 1996 dal Computer science & artificial intelligence Lab dell'Università di Palermo[9].

Un altro avanzato progetto di *visual retrieval* è il sistema VIPER (Visual Information Processing for Enhanced Retrieval), messo a punto nei laboratori dell'Università di Ginevra. Nell'interfaccia le ricerche possono essere impostate a partire dal *browsing* di una prima serie di immagini proposte dal sistema, ognuna delle quali è rappresentativa di una categoria e consente, se selezionata e inviata come dato di *query*, di continuare la ricerca con criteri propriamente visivi. È prevista anche la possibilità di interrogare il database partendo da un'immagine campione immessa dall'esterno o recuperata da un sito Web[10].

Infine, tra i programmi *storici*, si devono citare ImageQuery e Image Database Project dell'Università di Berkeley, risalenti nella prima formulazione al 1986, anche se poi riformulati intorno al 1990. Il progetto più importante è stato quello dell'interfaccia ImageQuery, nata con lo scopo di mettere a disposizione di tutti i dipartimenti dell'Università uno dei primi sistemi di differenti *database* di immagini digitali. L'UC Berkeley digital library project si è oggi evoluto in un'ampia serie di prospettive[11].

4 *Video retrieval*

Come tutti i sistemi informatizzati di *information retrieval*, anche quelli di *video retrieval* (recupero *content-based* di documenti video) sono solitamente composti da alcuni subsistemi tipici, che realizzano l'estrazione degli attributi dai dati, la modellizzazione e l'indicizzazione dei dati stessi, l'abbinamento tramite *query* e l'interazione con l'utente.

Fino a non molto tempo fa, larghezza di banda, spazio disco, memoria e potenze computazionali non erano in grado di supportare adeguatamente i dati non testuali. Negli ultimi anni tale situazione è radicalmente cambiata grazie ai progressi delle nuove tecnologie: con la sempre maggiore velocità delle CPU, più alte capacità di memoria ed efficienti metodi di compressione, i video digitali sono entrati nella nostra vita quotidiana, e il Web ha contribuito in modo notevole a favorire lo sviluppo di applicazioni per accedere ad archivi e collezioni *online*. L'elaborazione delle informazioni video è stata proprio una delle grandi sfide nello studio di *database* efficienti nel *retrieval*, poiché esse richiedono notevoli quantità di spazio di memoria e di potere computazionale, e presentano innumerevoli tipologie (dai programmi televisivi ai film, dalle registrazioni di conferenze a quelle di gruppi di lavoro, dai video *on demand* alle animazioni 3D).

Per il *video retrieval* vi è poi una ulteriore complicazione: la gestione dei dati, visuali, testuali e/o audio video - peraltro senza ancora definizione di standard univoci per la mancanza del necessario tempo di sedimentazione delle codifiche - è piuttosto complessa. Il recupero dei dati video condivide infatti alcune caratteristiche con il recupero dei dati delle immagini, per via della comunanza della loro natura visuale. Tuttavia, analogamente ai dati audio, i video sono anche dipendenti dal tempo e presentano in genere sia dati visivi che tracce audio. Queste caratteristiche in comune portano naturalmente ad applicare soluzioni proprie delle aree del *visual* e *audio retrieval* ai problemi del *video retrieval*. Ma tale strategia si dimostra solo talvolta effi-

9 JACOB è presentato all'indirizzo <www.csai.unipa.it/research/projects/jacob>.

10 Una demo di VIPER è all'indirizzo <viper.unige.ch>.

11 Una presentazione dell'intero UC Berkeley digital library project si trova all'indirizzo Web <elib.cs.berkeley.edu>.

cace, per le proprietà uniche dei video che richiedono soluzioni autonome e nuove per la classificazione, l'interrogazione e la presentazione dei dati.

4.1 Caratteristiche dei documenti video

Un flusso di dati video è una semplice sequenza di bit, che sarebbe priva di significato senza una qualche struttura sequenziale. Secondo un'architettura generale di sistema, i video acquisiti vengono infatti archiviati in un database e per essere resi disponibili per il *browsing* e/o il recupero in ambienti di rete devono essere analizzati, indicizzati e riorganizzati. Solo una volta che essi vengono adeguatamente predisposti, i dati video presentano la struttura adatta per un *browsing* non lineare e per un recupero *content-based* nell'ambito di grandi quantità di documenti. Per tale motivo, il *content-based retrieval* dei video richiede un'appropriata organizzazione dei dati, che possono essere distinti in quattro livelli o *layer*:
– *frame* (fotogrammi): unità di base del flusso di dati di un video;
– *shot*: un insieme registrato sequenzialmente di *frame*, che rappresentano un'azione continua nel tempo e nello spazio e derivano da una camera singola;
– *episodio*: serie di scene correlate. Una scena è una sequenza di *shot* che si focalizza sullo stesso punto o luogo di interesse. Una serie di scene correlate forma un episodio;
– *programma*: serie di episodi correlati.

Questa gerarchia fornisce in astratto una articolazione efficiente e flessibile di un video, caratterizzato anche da specifiche tipologie di struttura dell'immagine, del movimento degli oggetti, della camera e della transizione degli *shot*. Inoltre, nei video si riscontrano tipicamente modalità espressive multiple, poiché in genere sono accompagnati da colonne sonore, sottotitoli e/o titoli a schermo. Tale compresenza non produce ridondanza ma piuttosto sinergia: per questo motivo il *visual retrieval* e l'*audio retrieval* sono complementari all'analisi dei dati video e permettono di affrontare il problema della classificazione delle *query* in modi diversi (Fig. 7). Infine, data la loro natura dipendente dal tempo, i dati video condividono alcune delle difficoltà dei dati audio nelle modalità di rappresentazione dell'interfaccia e delle *query*.

L'indicizzazione delle parole chiave è un modo tradizionale di recuperare l'informazione ed è stato ampliamente applicato ai *database* video, così come l'applicazione di tecniche di analisi e di riconoscimento dei *pattern* delle immagini. Trattandosi però di processi computazionalmente onerosi e complessi, l'interesse dei ricercatori si è allora focalizzato sullo sviluppo di tecniche di indicizzazione che avessero la capacità di recuperare dati visivi in base ai loro contenuti, che fossero indipendenti dall'area tematica e che potessero essere automatizzati.

L'indicizzazione *content-based* dei video è basata sulla progettazione di un modello di dati, sulle relazioni tra attributi tramite mappature semantiche, sul riconoscimento dei dati audio tramite *wordspotting* o parole chiave selezionate dal riconoscimento automatico del parlato (cfr. il paragrafo successivo sull'*audio retrieval*). Un'altra strategia di classificazione è quella di utilizzare i *keyframe* (*frame* le cui immagini rappresentano una unità semantica di un flusso, come ad esempio una scena). Un problema cruciale con i sistemi di recupero di seconda generazione, in generale, rimane ancora il "salto" semantico tra sistema e utenti. Virtualmente tutti i sistemi proposti usano soltanto rappresentazioni significative di basso livello dei dati visuali, con una semantica limitata e facendo affidamento al *feedback* degli utenti. Negli ultimi anni, per ovviare al problema, è stato proposto un approccio basato sulla semiotica, che combina questi attributi di basso livello con il loro cambiamento nel tempo e con la narrazione del video (protagonisti, ruoli, azioni, relazioni), permettendo all'utente di identificare sia il ritmo della sequenza che la significanza.

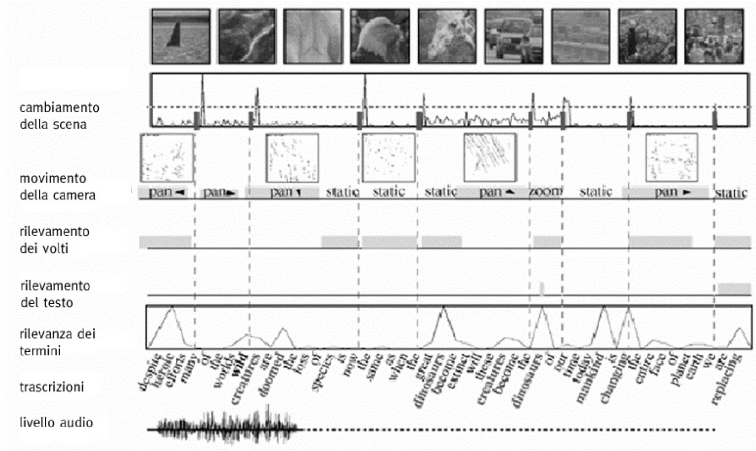

Fig. 6. Utilizzo di tecniche multiple per l'analisi e l'indicizzazione del flusso di dati di un video

4.2 Tecniche e applicativi

Come in generale per gli altri sistemi informativi, l'accesso a *database* di documenti video presenta due componenti principali - l'archiviazione e il recupero - e implica l'analisi dei contenuti e l'estrazione degli attributi, la modellazione del contenuto, l'indicizzazione e l'interrogazione.

Nel processo di archiviazione, i video vengono elaborati per estrarne gli attributi che descrivono il loro contenuto, che sono rappresentati, organizzati e conservati nel database. L'elaborazione dei dati può includere la decompressione, il potenziamento, il filtro, la normalizzazione, la segmentazione e l'identificazione dell'oggetto. L'*output* di questo stadio di elaborazione è tipicamente una collezione di oggetti e di regioni di interesse.

Nel passaggio successivo vengono estratti e rappresentati gli attributi dei contenuti (Fig. 8): nei video gli attributi spaziali sono per lo più generati usando le tecniche delle immagini fisse, mentre gli attributi temporali sono estratti basandosi sul movimento e/o sulle operazioni della camera in un determinato *shot.* Le tecniche di indicizzazione spaziale infatti trattano le sequenze video come immagini fisse, perdendo quindi la semantica contenuta nella sequenza e limitando le *query* degli utenti. Gli attributi temporali permettono invece all'utente di specificare nelle *query* l'esatta posizione e le traiettorie degli oggetti presenti in uno *shot*.

Come già indicato, il video presenta di per sé una gerarchia di unità con *frame* singoli al livello di base e segmenti di più alto livello come *shot*, scene ed episodi. Un compito importante nell'analisi dei contenuti video è la rilevazione dei confini dei diversi segmenti, che riflettono le convenzioni editoriali tradizionali. Ma gli indici dei dati potrebbero essere rappresentati in modo approssimativo, possedere attributi multipli interrelati e non avere un ordine intrinseco chiaro. Per tale motivo non possono essere utilizzati i tradizionali approcci di indicizzazione, ma occorrono delle strutture più flessibili. Infine, è necessario che il *database* fornisca all'utente una rappresentazione concettuale dei dati, che ne supporti viste multiple e ne assicuri la consistenza.

La segmentazione costituisce perciò un passaggio cruciale per un sistema di *video retrieval*. Nel processo di indicizzazione vengono utilizzati degli algoritmi per rilevare i cambiamenti della scena e per suddividere il flusso dei dati video in segmenti signi-

ficativi più facilmente gestibili e utilizzabili come unità di base per l'indicizzazione, per l'appunto gli *shot* (una sequenza di *frame* che hanno gli stessi contenuti). Una volta che gli *shot* sono stati identificati, vengono estratti i *keyframe* per permettere la ricerca per ogni segmento ed elaborare la *query* sui *frame* selezionati. Questa modalità di ricerca tuttavia non sempre garantisce dei risultati corretti nella fase di recupero, per via della difficoltà di interpretazione dovuta alla complessità dei dati video.

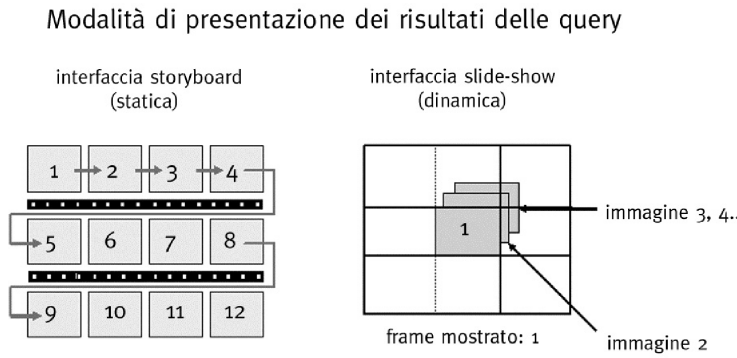

Fig. 7. Tecniche di rappresentazione per i surrogati statici
e dinamici di documenti audio: modalità storyboard e slideshow

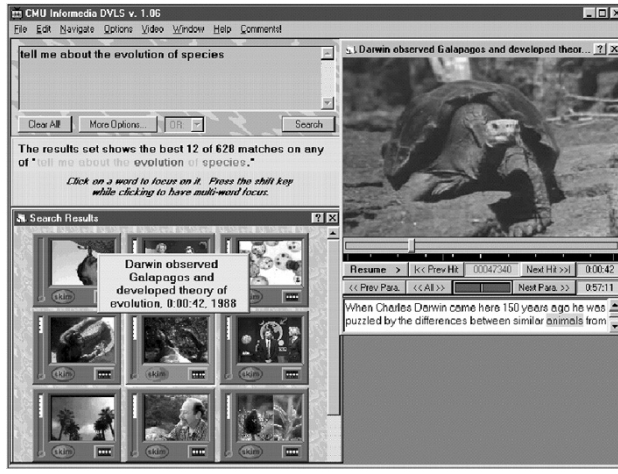

Fig. 8. Progetto *Informedia* (Carnegie Mellon University):
l'applicazione di *news-on-demand* indicizza le notizie della TV
e della radio e permette agli utenti di recuperare i *frame* tramite il contenuto

Le tecniche per il rilevamento dei cambiamenti degli *shot* includono il confronto diretto dei pixel o dell'istogramma, l'analisi degli attributi compressi e il riconoscimento del testo, dei sottotitoli e dei titoli a schermo per l'indicizzazione video. Per l'annotazione dei contenuti dei video, le trascrizioni del parlato sono state ad esem-

pio utilizzate nel progetto Informedia della Carnegie Mellon University[12] come una fonte preziosa di informazioni (Fig. 9).

La precisione del riconoscimento degli oggetti nelle immagini di una scena non è molto alta, poiché lo scopo principale dei metodi di *video retrieval* è il recupero di immagini complesse che contengono più oggetti. Dopo la segmentazione iniziale, nei video vi sono due fonti di informazioni che possono essere utilizzati per rilevare e tracciare gli oggetti: attributi visivi (come il colore della pelle, la *texture* e la forma) e informazioni sul movimento. Ricerche nell'area dell'*image processing* hanno introdotto dei subsistemi di riconoscimento degli oggetti per estrarre automaticamente gli attributi dei contenuti visuali a partire dalla distribuzione dei pixel. I subsistemi di analisi dell'immagine supportano l'estrazione automatica delle caratteristiche percettive come il colore e la *texture*, o primitive semantiche come la forma e le relazioni spaziali. Nel rilevamento del movimento, gli oggetti vengono ipotizzati come nell'*image retrieval* (segmenti basati sul colore e sulla *texture*), tramite l'esame dei cambiamenti dei pixel da un *frame* all'altro e la classificazione dei *pattern* risultanti dal movimento (traslazione, rotazione, unione o divisione). Sono in corso di sviluppo sistemi per rilevare particolari movimenti (come le entrate e le uscite da una scena e il posizionamento/rimozione di oggetti), espressioni facciali e i diversi *speaker*. Le tecniche di *machine learning* permettono di aumentare la precisione nei risultati, ma l'addestramento del computer richiede grandi quantità di dati visionati e un notevole dispendio computazionale.

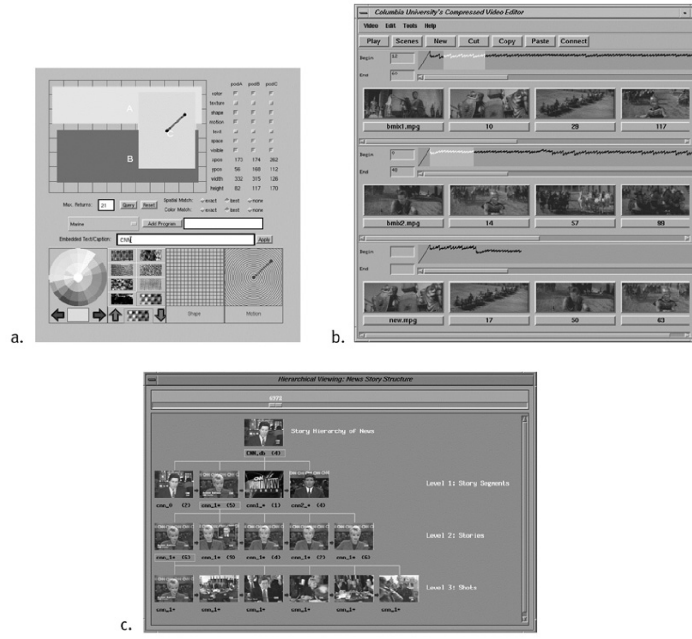

Fig. 9. VisualSEEK: differenti livelli di attributi visivi.
a. attributi di livello base (*texture*, colore, movimento);
b. informazioni derivanti dalla conoscenza dell'area tematica;
c. astrazione di informazioni di alto livello

12 Cfr. <www.informedia.cs.cmu.edu/>.

Per permettere l'accesso al *database* da parte dell'utente, la natura temporale e la grandi dimensioni dei file dei video richiedono l'implementazione di funzioni specifiche di *query* e *browsing*. Nel processo di interrogazione il modello strutturale di un video permette un insieme molto ricco di *query* spaziali (come "parallelo a" e "sotto") e temporali (come "segue", "contiene" e "transizione") grazie ad un Hierarchical Temporal Language (HTL). La semantica di questo linguaggio viene progettata per un recupero basato sulla similarità: la *query content-based* implica il rilevamento e il tracciamento tra gli oggetti in movimento nel video e quelli nell'esempio fornito/selezionato dall'utente.

Molti sistemi sviluppati in ambito commerciale o di ricerca forniscono indicizzazione e *query* automatica basata su attribuiti visuali come il colore o la *texture*: ad esempio VisualSEEk[13] e WebSEEk[14] della Columbia University, AT&TV della Cambridge University e della AT&T [15], i sistemi della Virage[16], il sistema *ECHO* del CNR-Pisa[17], il progetto Cultural Digital Library Indexing our Heritage (CLIOH) della IUPUI University[18]. VisualSEEk (Fig. 10) e WebSEEk, già citati nel precedente paragrafo sul *visual retrieval*, sono dei sistemi di ricerca e recupero delle immagini che permettono *query* di attributi locali e il perfezionamento degli istogrammi per il *feedback*, usando uno strumento *Web-based*. WebSEEk, in particolare, costruisce molteplici indici per le immagini e i video, basandosi su attributi visuali quali il colore e attributi non visuali come le parole chiave assegnate alle tipologie di immagini e di video.

Infine, un criterio importante per la performance di un sistema di *video retrieval*, data la presenza dei subsistemi sopra indicati, è la qualità del servizio. Poiché la maggior parte degli utenti accedono al sistema tramite una qualche forma di *network*, bisogna assicurare la continuità e la sincronizzazione dei flussi multimediali tenendo conto delle limitazioni dei subsistemi: per tale motivo sono state proposte e implementate varie tecniche di *buffering* e *scheduling*.

5 *Audio retrieval*

I sistemi informatizzati di recupero di documenti sonori (*audio retrieval*) possono sommariamente essere suddivisi in due categorie: sistemi di riconoscimento automatico del parlato (*automatic speech recognition*) e sistemi di recupero delle informazioni sonore in generale, dalla musica (*music information retrieval*) ai rumori (incluso il parlato).

L'*audio retrieval* rappresenta un'area di ricerca relativamente recente ma i problemi che essa affronta sono familiari a tutti coloro che devono richiamare un suono o una musica selezionandoli tra diversi esemplari. Stimolati dalla sempre crescente disponibilità di grandi collezioni di documenti audio in formato digitale e dalle possibilità offerte dalle tecnologie di rete, negli ultimi anni alcuni ricercatori hanno sviluppato tecnologie di elaborazione del parlato e della musica, creato modelli della percezione uditiva e inventato nuovi modi di analizzare, rappresentare e sintetizzare i segnali audio. Queste ricerche hanno trovato un fer-

13 Cfr. <www.ctr.columbia.edu/~jrsmith/VisualSEEk/VisualSEEk.html>.

14 Cfr. <www.ctr.columbia.edu/webseek/>.

15 Cfr. <www.uk.research.att.com/dart/attv/index.html>.

16 Cfr. <www.virage.com>.

17 Cfr. < http://pc-erato2.iei.pi.cnr.it/echo/ >.

18 Cfr. < http://clioh.informatics.iupui.edu/>.

tile terreno d'applicazione sia nei settori dello spettacolo, dell'intrattenimento e della comunicazione (distribuzione della musica, librerie di suoni digitali per film, animazioni, videogiochi, mezzi di comunicazione televisivi, radio e digitali), che in varie aree disciplinari (interazione uomo-computer, interfacce grafiche e acustiche, bioacustica, psicoacustica, percezione musicale, etnomusicologia, musica computerizzata, realtà virtuale).

Gli approcci dei sistemi di recupero testuale (localizzazione dei documenti di testo desiderati usando una *query* di ricerca con parole chiave) e visuale (standard, classificazione degli attributi e valutazione) sono in parte ma non del tutto applicabili anche ai sistemi di *audio retrieval*. Ciò dipende dalla specifica natura dei documenti audio, che dipendono dalla dimensione temporale e che sono (ovviamente) di tipo acustico e non di tipo visivo. Le proprietà e le caratteristiche dell'audio sono suddivisibili in due categorie generali, che permettono di distinguere e classificare il parlato, la musica e i rumori:

1. proprietà e caratteristiche relative alla dimensione temporale: i segnali audio possono essere espressi come ampiezze che variano nel tempo e che presentano attributi come l'altezza tonale (*pitch*) e il *silence ratio*;
2. proprietà e caratteristiche relative alla frequenza: la frequenza viene definita come pari a 1/tempo, e lo spettro sonoro è l'ampiezza (la quantità di energia) che varia con la frequenza. Tra le caratteristiche vi sono ad esempio la larghezza di banda, la gamma di frequenza sonora, la differenza tra la frequenza minima e massima, l'armonica, lo spettrogramma.

La dimensione acustica richiede approcci creativi per risolvere le questioni relative all'interrogazione e al recupero dei dati, poiché è possibile che l'audio non contenga parole e in ogni caso presenta il problema della linearità. La dimensione temporale pone dei problemi di presentazione specifici, che i sistemi di recupero testuale, visuale e video non hanno o presentano in misura molto minore. Inoltre, per il recupero e la navigazione i documenti audio coprono complessivamente un numero di situazioni molto diverse, in cui – come accade in generale per l'i*nformation retrieval* – è necessario specificare di volta in volta la natura e la scala dei componenti che vengono cercati: recupero e navigazione tra segmenti di uno stesso documento sonoro; indicizzazione e recupero di brevi suoni individuali in basi dati audio; recupero di documenti sonori in un insieme relativamente consistente di documenti dello stesso tipo; recupero e navigazione tra segmenti di video che usano informazioni sonore di tipo diverso; recupero tramite selezione dei termini in basi dati di parlato o di video, con indicizzazione del parlato.

5.1 *Automatic speech recognition*

Grazie ai recenti sviluppi negli algoritmi di ricerca, alla potenza computazionale dei processori e al ridotto costo delle memorie, la tecnologia di riconoscimento automatico del parlato sta rapidamente passando dai laboratori di ricerca alla nostra vita di ogni giorno, sotto forma di applicazioni commerciali (dall'interazione con i sistemi automatizzati telefonici al controllo di programmi informatici) che trattano il parlato nello stesso modo in cui viene attualmente trattato il testo nell'i*nformation retrieval* tradizionale, e che permettono un riconoscimento in *real-time*. Le registrazioni del parlato vengono pre-elaborate traducendole in forma testuale (ad esempio trascrizioni dell'audio) oppure operando direttamente sulle registrazioni audio. Le trascrizioni in genere non sono ancora perfette e per tale motivo sono stati sviluppati degli strumenti informatici che ne migliorano la qualità, come ad esempio rendere automatiche le maiuscole nel

caso di testi tutti in minuscolo o segmentare il testo, dato che i sistemi di *automatic speech recognition* (ASR) sono in grado di riconoscere con maggiore facilità i termini con più caratteri.

Da quello fonologico a quello pragmatico, vi sono vari livelli di analisi linguistica, che riflettono una crescente complessità e difficoltà del linguaggio. Molti motori di ricerca utilizzano sistemi di i*nformation retrieval* basati sulle modalità elementari di ricerca linguistica, ma la ricerca per termini presenta dei difetti intrinseci: le parole vengono cercate al di fuori del contesto, e la terminologia utilizzata è spesso ben lungi dall'essere precisa, presentando molteplici sfumature e mancanza di accuratezza.

Soprattutto, più l'unità analizzata diventa grande (ad esempio dai morfemi alla parole, dalle frasi ai paragrafi, all'intero documento), meno precisi risultano i fenomeni del linguaggio e maggiore la possibilità di scelta e le variabili. Inoltre i livelli più alti presuppongono la comprensione sia dei livelli inferiori, sia delle teorie impiegate per spiegare i dati, che devono necessariamente spostarsi nel campo della psicologia cognitiva e dell'intelligenza artificiale. Come risultato, la mancanza di mediazione semantica influisce in modo ancora determinante sulle prestazioni dei sistemi di i*nformation retrieval* testuali e su quelli che analizzano il parlato. Lo svantaggio principale di questi sistemi è la loro limitata accuratezza: sebbene i migliori sistemi di ricognizione continua riescano a raggiungere il 90% di accuratezza nelle parole su aree tematiche circoscritte, altri sistemi simili riescono a raggiungere solo un 50-60% di accuratezza in situazioni del mondo reale come le conversazioni telefoniche e i telegiornali. Nonostante questo, però, i risultati delle trascrizioni possono essere molto utili per il recupero dei dati del parlato, con applicazioni che vanno dall'utilizzo durante le riunioni di lavoro alle teleconferenze e alla trascrizione del parlato in sedi istituzionali.

In questa direzione il Commonwealth Scientific & Industrial Research Organization (CSIRO) australiano ha definito un approccio basato sulla tecnologia Annodex, che consente l'annotazione e l'indicizzazione automatica dei file audio e video sul Web[19]. Tale sistema permette sia di cercare che di linkare i file (rendendoli quindi direttamente identificabili dai motori di ricerca), che di spostarsi al loro interno, con implicazioni sostanziali per l'infrastruttura ipertestuale e di *browsing* della Rete (Fig. 11).

Quasi tutti i sistemi di ASR in uso sono basati su rappresentazioni statistiche di eventi del parlato (come un parola) in cui i parametri del modello vengono addestrati su una base dati di grandi dimensioni (come il *corpus* SWITCHBOARD[20]): dato un insieme addestrato di Hidden Markov Model (HMM), si applica un algoritmo d'efficienza per trovare la sequenza del modello più probabile e perciò più corretto (combinazioni di parole riconosciute) in dati di parlato sconosciuto. Tuttavia se una parola non è presente nel vocabolario fonetico essa non verrà riconosciuta, potrebbe non essere possibile trovare un testo esemplificativamente sufficiente per la modelizzazione e, infine, nonostante nuovi algoritmi tali sistemi sono ancora abbastanza onerosi in termini computazionali e di memoria.

19 Cfr. <www.annodex.net/>.

20 Cfr. <www.isip.msstate.edu/projects/switchboard/ >.

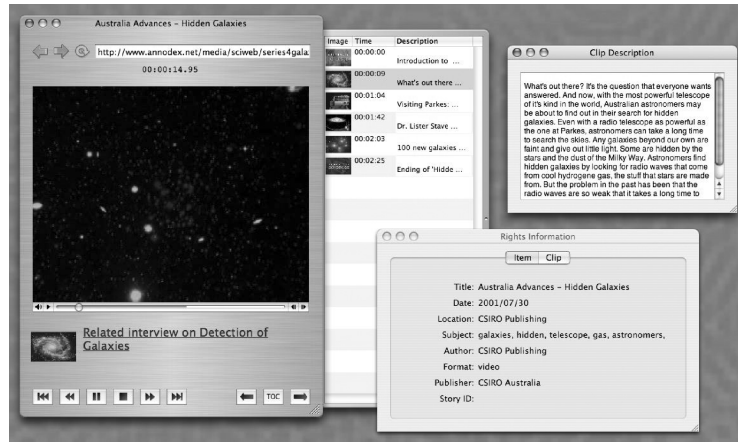

Fig. 10. *Screenshot* della demo del *browser* CMWeb sviluppato dal gruppo di ricerca CMWeb della CSIRO utilizzando la tecnologia Annodex

Per ovviare a tali inconvenienti è possibile applicare vari metodi. Il *keyword-spotting*, ovvero la rilevazione automatica di alcune parole o frasi nel parlato naturale, rende l'analisi computazionalmente meno onerosa e sufficientemente flessibile per gestire il parlato naturale del mondo reale. Il metodo *sub-word* utilizza invece unità tipicamente più piccole delle parole intere (di solito basate uno specifico dizionario fonetico che indica come concatenare le unità per formare un termine), riducendo notevolmente lo spazio di ricerca. Uno svantaggio, tuttavia, è che con unità di dimensioni ridotte diminuisce anche l'accuratezza del riconoscimento. I metodi di identificazione degli *speaker* (detti anche *speaker ID*) (Fig. 12) e dei linguaggi, invece, usano modelli addestrati all'identificazione della differenza tra voci e lingue piuttosto che alla determinazione di ciò che viene detto, utilizzando basi dati come il *corpus* TIMIT[21], che contiene registrazioni di 630 persone in otto dei maggiori dialetti anglo-americani.

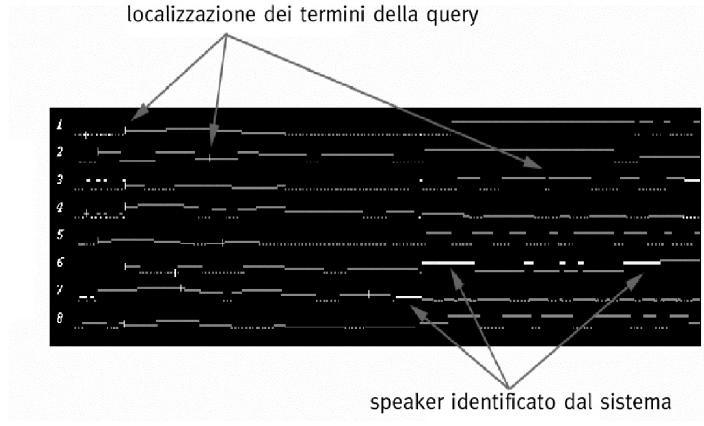

Fig. 11. Esempio di sistema di identificazione dello *speaker* in 8 registrazioni audio

21 Cfr. <www.ldc.upenn.edu/Catalog/docs/TIMIT.html>.

Tra gli esempi di applicazioni più significativi, vi sono il già citato progetto Informedia della Carnegie Mellon University e il sistema AT&TV per la ricerca in telegiornali e programmi radio tramite indicizzazione in *realtime*, sviluppato congiuntamente dalla Cambridge University e dalla AT&T, entrambi citati nel precedente paragrafo sul *video retrieval*. Il sistema di Informedia sfrutta le ricerca sui *browser* video della FX-PAL e il Virage Audio logger™ (usando tecnologie della IBM e di Muscle Fish), offrendo una impressionante combinazione di analisi video e audio e di tecniche di recupero testuale[22]. Si segnala inoltre il sistema commerciale *Speechbot - Audio Search using Speech Recognition* della Hewlett Packard[23], che effettua le ricerche attraverso oltre 10.000 ore di programmi radio indicizzati. Tra i sistemi realizzati in ambito accademico negli anni passati vi sono invece il Radio news retrieval del Eidgenössische technische Hochschule (ETH) di Zurigo e lo SpeechSkimmer del Massachusetts Institute of Technology (MIT) e il progetto Video Mail Retrieval (VRM) della Cambridge University in collaborazione con Olivetti research laboratory[24].

5.2 *Music information retrieval*

L'universo possibile dell'audio è naturalmente molto più ampio del solo parlato, e la musica rappresenta una classe di documenti incredibilmente ampia e varia, dato che la definizione stessa di "musica" non è universalmente condivisa. Considerando l'ampia gamma di suoni che le persone vorrebbero archiviare o classificare, dai generi musicali agli effetti sonori ai versi degli animali ai campioni dei sintetizzatori, è chiaro che i metodi basati sul riconoscimento del parlato sono incredibilmente inadeguati.

A complicare questa operazione c'è il fatto che la musica può essere monofonica o polifonica, e che suoni diversi possono accadere in concomitanza. Un requisito fondamentale per poter utilizzare i dati audio è quello di poter rappresentare la loro struttura temporale. I linguaggi di rappresentazione della musica sono utili per molteplici impieghi:

– scambio di file indipendente dalla piattaforma utilizzata (software musicali, Web);
– accesso intelligente e dinamico a musica strutturata (biblioteche, didattica);
– utilizzo della funzionalità di riconoscimento musicale per fornire accesso ad una massa critica di file musicali (editoria, distribuzione, vendita, esecuzione di documenti musicali);
– processi collaborativi di creazione e lavoro e sincronizzazione di *media* dello stesso tipo o di tipi diversi basati sul tempo (industria di produzione video).

A partire all'incirca dal XIX secolo, vi sono stati diversi tentativi di linguaggi di rappresentazione della musica (come mostrato nella tabella 1), ma nessuno è stato finora universalmente accettato. Poiché è più facile lavorare con le rappresentazioni simboliche ed esse richiedono minore elaborazione, molti studi sul *music information retrieval* si sono focalizzati ad esempio sulle *music instrument digital interfaces* (MIDI), uno standard è stato inventato per catturare i gesti di chi suona una tastiera, e permette di fornire una visione semplificata del contenuto di altezza tonale e di durata di un documento musicale. Tuttavia, la maggior parte della musica è digitalmente disponibile in forma di segnali audio grezzi, non strutturati e monolitici, che richiedono un ampio

22 Il progetto *Informedia* e numerose relative pubblicazioni sono consultabili a <www.informedia.cs.cmu.edu/>.

23 Cfr. <http://speechbot.research.compaq.com/>.

24 Cfr. <http://mi.eng.cam.ac.uk/research/Projects/vmr/vmr.html >.

e sofisticato utilizzo dell'elaborazione dei segnali e di algoritmi di *machine learnig*. Lo standard MPEG-7, offrendo un formato strutturato di dati, promette certamente sviluppi significativi in quest'ambito[25], come ha illustrato Michael Casey[26] e come è stato mostrato nel software AudioID sviluppato dal Fraunhofer IIS[27].

L'analisi audio si riferisce in generale alle tecniche di pre-elaborazione, che aiutano il computer a "capire" il suono in modo simile a quello di un essere umano. In particolare, un problema generale nell'analisi di file audio è quello della discriminazione del parlato dalla musica non vocale o da altri suoni: in generale, è importante non sprecare risorse preziose per la classificazione e il recupero dei dati cercando di effettuare un riconoscimento del parlato su brani musicali, sul silenzio o su altro tipo di suono.

Alcuni esempi di analisi audio sono la classificazione, la segmentazione, il *retrieval* e il *clustering*. Tipicamente, l'analisi del suono si basa sul calcolo dei vettori delle caratteristiche che descrivono il contenuto spettrale di un suono. In genere viene utilizzato il riconoscimento statistico dei *pattern*, basato sull'estrazione delle caratteristiche spettrali del file audio. I primi studi in quest'area si sono concentrati sulla similarità melodica in raccolte monofoniche, basandosi sull'elaborazione del segnale (*signal processing*) per estrarre un insieme di caratteristiche acustiche per ogni suono. Più recentemente si è cominciato a valutare anche gli aspetti cognitivi e percettivi delle melodie e sono stati sviluppati dei metodi che affrontano la complessità della musica polifonica, ad esempio applicando l'indicizzazione di segmenti musicali in base all'osservazione del comportamento degli utenti.

Gli utenti delle tecnologie di *music information retrieval* comprendono varie tipologie, da chi semplicemente ascolta la musica a compositori, musicisti, musicologi, analisti forensici. Le persone in genere ricercano altri "sistemi di *retrieval* umani" (ad esempio bibliotecari di discipline musicali, radio DJ e commessi di negozi di musica) includendo le informazioni contestuali a disposizione e canticchiando parte di una melodia. Per tale ragione, il riconoscimento automatico di una melodia è un problema che ha catalizzato molte ricerche. I sistemi di *music information retrieval* spesso adottano un approccio *content-based* tramite una *query-by-example*, poiché può essere non poco frustrante formulare una *query* testuale per ottenere dati audio.

L'immissione di una *query* può avvenire tramite una tastiera MIDI oppure canticchiando (*query-by-humming*) nel microfono del proprio computer. Dopo che il sistema ha accettato la *query* acustica e l'ha codificata in formato digitale, vi sono diversi modi di eseguire l'abbinamento con dati audio simili a quelli immessi. In generale, usando astrazioni che ammettono errori come l'analisi di frequenza o il *pitch contour*, i sistemi di *audio retrieval* possono trasformare il problema dell'abbinamento audio con i ben noti problemi di calcolo delle distanze di *edit* o di abbinamento delle stringhe. Tuttavia, la struttura multidimensionale e spesso complessa dei dati rende complessa sia la formulazione delle *query* che l'abbinamento con i dati archiviati; in secondo luogo, vi è spesso un considere-

25 Cfr. <www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>.

26 Michael Casey, *General audio information retrieval*, in *MultiMedia Information Retrieval: metodologie ed esperienze internazionali di content-based retrieval per l' informazione e la documentazione*, Roma: AIDA, 2004, p. 369-384.

27 Cfr. <www.idmt.fraunhofer.de/projekte_themen/index.htm>.

vole grado di soggettività, incertezza o accuratezza nella *query* e/o nei dati, derivante dai metodi utilizzati per le *query* (come il canticchiare) o per acquisire i dati musicali (come la trascrizione automatizzata dell'esecuzione musicale) oppure dal semplice errore umano nell'inserire i dati. Inoltre, un approccio basato sull'informazione delle note non permette di ricercare tutti i tipi di musica: ad esempio alcune musiche create nel XX secolo non sono basate sulle note, e consistono di suoni invece che di melodie, e vi sono diversi tipi di scale musicali. Per cercare di ovviare a tali difficoltà si usano spesso degli algoritmi di abbinamento approssimativo.

Alcuni recenti progetti hanno tentato di abbinare più caratteristiche invece della sola informazione melodica. Ad esempio in MuscleFish[28] la scelta viene effettuata in base alla similarità con delle caratteristiche psicoacustiche predefinite, e la classificazione e il recupero dei dati sono basati sulle tecniche tradizionali di analisi dei dati. Nel progetto Recognition and Analysis of Audio (RAA)[29] questo processo viene applicato nell'ambito della gestione dei diritti di proprietà intellettuale, della vendita di musica *online* e degli archivi musicali.

Una volta che l'utente ha inviato una *query* e il sistema ha determinato il numero di abbinamenti possibili nel database, il passo logico successivo è quello di permettere il *browsing*, la navigazione e la visualizzazione dei dati audio. In questa operazione la natura dipendente dal tempo dei dati audio pone il problema della rappresentazione. Per i *media* che non sono dipendenti dal tempo, come il testo e le immagini, i dati sono statici e possono essere visualizzati senza problema. Ma in quelli come l'audio si presentano varie difficoltà relative alla trasposizione o meno di tutte le combinazioni alla stessa chiave per facilitare i confronti, all'utilizzo o meno una rappresentazione visiva per presentare l'audio (permettendo l'equivalente di una veloce scansione attraverso vari abbinamenti per identificare l'abbinamento appropriato), alla rappresentazione di *query* abbinate nel loro contesto e alla creazione di riassunti del flusso dei dati audio (*audio thumbnailing*) per rendere più veloci le valutazioni di rilevanza da parte dell'utente. Ad esempio, mentre non risulta problematico mostrare contemporaneamente diversi *keyframe* di vari video, nel caso dell'audio suonare simultaneamente 15 clip musicali (ovvero 15 *query* di abbinamento rappresentate tutte insieme) non risulterà molto utile per chi effettua la ricerca.

Vi sono molti strumenti per analisi e la sintesi del suono: tipicamente operano su un singolo file audio e si servono di una grafica 2D per la visualizzazione, che però permette un'interazione limitata. Più recentemente, sono stati sviluppati anche degli strumenti per la visualizzazione grafica 3D in *real-time* in grandi collezioni di file audio. Tra le ricerche più significative, si segnala in particolare MARSYAS[30], un *framework* software per applicazioni di *computer audition* realizzato dalla Princeton University: il progetto presenta una serie di *browser* implementati in un'interfaccia grafica (Fig. 13), che funziona su computer commerciali senza bisogno di hardware specifico per la grafica 3D o l'elaborazione di segnali audio.

28 Cfr. <www.musclefish.com> <www.musclefish.com/cbrdemo.html>.

29 Cfr. <http://raa.joanneum.ac.at/start.html>. Altri progetti simili sono IMPRIMATUR <www.imprimatur.alcs.co.uk/central>, SDMI (Secure Digital Music Initiative) <www.sdmi.org> e VERDI (Very Extensive Rights Data Information) <www.verdi-project.com>.

30 Il software è gratuito è può essere scaricato dal sito <http://www.cs.princeton.edu/~gtzan/marsyas.html>.

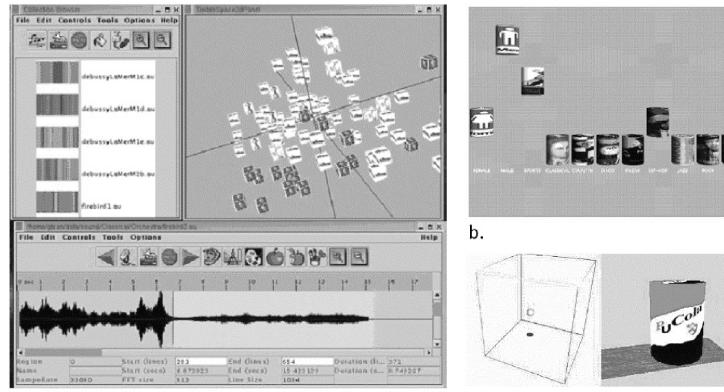

Fig. 12. Alcuni subsistemi del *framework* MARSYAS:
a. *TimbreGram*;
b. *GenreGram Browser* per la visualizzazione dei generi musicali;
c. *TimbreBall Browser* per la visualizzazione animata *real-time* dei vettori degli attributi

6 Conclusioni

Il *MultiMedia Information Retrieval,* naturale e irrinunciabile sviluppo dei principi dell'*information retrieval*, costituisce un *metasistema* dove la formulazione di richiesta informativa non è costretta entro i limiti della lingua, ma può essere inviata al dispositivo di input così come è spontaneamente prodotta dall'utente, in caratteri immediatamente visivi, sonori o audiovisivi, nonchè testuali nei casi specifici, e così come prodotta può essere dal sistema afferrata, trattata e soddisfatta: attraverso testi, linee, colori, forme, strutture, suoni, frequenze, movimenti, flussi e così via.

Il metodo del MMIR è fortemente innovativo, specializzato per un efficiente trattamento di ogni genere di informazione multimediale digitale. Metodo basato su una tecnologia di archiviazione e recupero definita *content-based* proprio perché tratta direttamente il *contenuto oggettivo* dei documenti, in opposizione ai tradizionali sistemi di indicizzazione e ricerca basati solo su *termini descrittori* di tale contenuto multimediale, detti *term-based*.

I due sistemi, però, possono e devono essere *integrati*: l'interrogazione *term-based* può, intanto, essere un ottimo metodo preliminare per selezionare una parte della grande quantità di documenti di un archivio, e per centrare la ricerca in base a dati quali gli ambiti d'appartenenza dei documenti, le tipologie, le classi, i titoli, gli autori; quindi, essa può essere un sistema finale di ripulitura dall'inevitabile *rumore* specifico di un'interrogazione *content-based*.

Il MMIR, in definitiva, realizza la possibilità di ricercare i testi, le immagini, i video e i suoni caso per caso tramite i più appropriati mezzi dei linguaggi testuale, visivo, audiovisivo e sonoro, quali la somiglianza, l'approssimazione e i rapporti di misure e valori, utilizzando come chiavi di recupero figure, strutture, suoni, testi, forme e colori, anche in un'unica schermata di ricerca nella formulazione di una sola complessa *query* multimediale.

Come si è potuto vedere, i subsistemi che compongono il MMIR condividono alcune problematiche comuni, ma anche notevoli possibilità di utilizzo e di sviluppi concreti.

Nel caso del *visual retrieval*, la sua specifica tecnologia si identifica nelle linee del sistema generale assimilando i presupposti del *content-based retrieval* e applicandoli alle immagini

fisse. È necessario anche in questo caso ribadire che un buon livello di precisione nel recupero dei documenti visivi non si può raggiungere senza utilizzare in combinazione, e mutua integrazione, tecniche e tecnologie di ricerca basate sia sulla definizione dei concetti, tramite termini controllati, sia sulla rappresentazione del contenuto, attraverso elementi visivi. I due procedimenti possono e devono operare sempre in armonia e in interazione costante, realizzando anche all'interno dello specifico settore del *visual retrieval* quella prospettiva di organica interrelazione che è propria del MMIR e degli altri ambiti in esso integrati.

Per quanto riguarda il *video retrieval*, una ridotta percentuale dei sistemi creati è stata sottoposta a verifiche e i test hanno spesso prodotto risultati mediocri. Inoltre, vi sono pochi sistemi di *benchmarking* che si rifanno ad uno standard comune, soprattutto perché i sistemi che analizzano e permettono le *query* di tutti gli attributi dei documenti multimediali sono complessi da valutare (la *performance* è relativa non soltanto al compito da eseguire ma anche alla modalità di interazione, alle caratteristiche dell'hardware che utilizza l'utente e alla scelta degli attributi). Altri elementi problematici sono poi costituti dal numero di sistemi prototipali esistenti, dalla natura dinamica della maggior parte dei sistemi e dalla formazione degli utenti al loro utilizzo. Gli attuali sistemi di *video retrieval* sono il risultato di ricerche combinate in vari campi: una migliore collaborazione tra le aree della *computer vision*, del *database management* e delle tecniche di interfaccia utente fornirà sistemi di recupero dei video e delle immagini più efficaci ed efficienti. Migliori tecniche di compressione e di tracciamento dei documenti multimediali incrementeranno l'accessibilità ai dati, mentre perfezionate tecniche di visualizzazione 2D e 3D e interfacce utente generiche e personalizzabili consentiranno all'utente di avere una visione d'insieme dei documenti di una collezione, ed essere così aiutato nella ricerca e nel *retrieval*. Anche la possibilità di *query* sonore aggiungerà una dimensione finora poco esplorata al *video retrieval*.

Le tecnologie dell'*audio retrieval*, infine, stanno maturando velocemente, e presto permetteranno applicazioni robuste per un largo e sistematico impiego, ma si devono ancora trovare soluzioni a numerose problematiche (ad esempio scalabilità su collezioni di grandi collezioni sia di soli dati audio che miste; capacità di gestire efficacemente sia la musica monofonica che quella polifonica e di lavorare con tutti i diversi tipi di musica; usabilità e interazione uomo-computer; progettazione di interfacce di *front-end* in grado di aiutare adeguatamente gli utenti con formati e azioni multipli; implementazione degli standard di comunicazione e dei formati dei dati audio).

Tuttavia, come è stato scritto, probabilmente nell'ambito documentale «music will constitute a "killer app" for digital libraries»[31].

Infine, *last but not least*, si segnalano le ricerche di Eric Paquet, *research scientist* del gruppo di Visual information technology del National Research Council canadese, che ha introdotto un nuovo approccio teorico per il riconoscimento di oggetti multimensionali dinamici. Si tratta di una proposta innovativa nell'ambito del *content-based retrilval*: il metodo di Paquet[32] è generalizzabile a quasi ogni tipologia di oggetto e può essere applicato in vari ambiti (in particolare in quello medicale), per la tomografia assiale computerizzata (TAC), la visione multispettrale e gli oggetti deformabili.

31 David Bainbridge [*et al.*], *Towards a digital library of popular music*, in: *Proceedings of the fourth ACM conference on digital libraries, Berkeley (CA), August 11-14, 1999,* New York: ACM Press, 1999, p. 169.

32 Paquet si occupa di *content based management* di oggetti multimediali e multidiemsionali, *database* antropometricim gestione e visualizzazione delle informazioni di siti storici e archeologici virtualizzati. Cfr. <www.cleopatra.ncr.ca/Nefertiti/EricPaquet/index.html>.

RIFERIMENTI BIBLIOGRAFICI

Allan James. *Perspectives on information retrieval and speech*, in: *Information retrieval techniques for speech applications* [based on the workshop *Information retrieval techniques for speech application*, held as part of the 24th Annual international ACL SIGIR conference on research and development information retrieval, New Orleans, september 2001], Anni R. Coden , Eric W. Brown, Savitha Srinivasan (eds). Berlin: Springer, 2002, p. 1-10. (Lecture notes in computer science, 2273).

Besser Howard. *Image and multimedia database resources*, 2000, <sunsite.berkeley.edu/Imaging/Databases>

Del Bimbo Alberto. *Visual information retrieval.* San Francisco: Kaufmann, 1999.

Image and video databases: visual browsing, querying and retrieval, Alberto Del Bimbo (ed.). «Journal of visual languages and computing», 7 (1996), n. 4 (special issue).

Dimitrova Nevenka – Golshani Forouzan. *Video and image content representation and retrieval*, in: *Handbook of multimedia information management.* William I. Grosky, Ramesh Jain, Rajiv Meehrotra (eds.). Upper Saddler River (NJ): Prentice Hall, 1997, p. 95-138.

Du-Seok Jin – Lee Jeong-Jae – Chang Jaeewoo I. *A structure- and content-based multimedia information retrieval system for XML documents*, in: *Challenges of information technology management in the 21st century: Information resources management association international conference, Anchorage (AK), May 21-24, 2000.* Medi Khosrowpour (ed.). Hershey (PA): Idea Group publishing, 2000, p. 342-350.

Downie J. Stephen. *Access to music information: the state of the art.* «Bulletin of the American Society for Information Science», 26 (2000), n. 5, <http://www.asis.org/Bulletin/June-00/downie.html>

Enser Peter G.B. *Pictorial information retrieval (Progress in documentation).* «Journal of documentation», 51 (1995), n. 2, p. 126-170.

Finn Robert. *Query by image content.* «IBM Journal of research and development», 40 (1996), n. 3, p. 22-32.

Foote Jonathan T. *Content-based retrieval of music and audio,* in: *Proceedings of the SPIE, Multimedia storage and archiving systems II, 3-4 November 1997, Dallas.* Chung-Chieh J. Kuo, ShihFu Chang, Venkat N. Gudivada (eds.). Bellingham (WA) : SPIE, c1997, p. 138-147. (Proceedings of the SPIE, 3229) <http://svrwww.eng-cam.ac.uk/~jtf/papers/spie97-abs.html>

Grosky William I. *Managing multimedia information in database systems.* «Communications of the ACM», 40 (1997), n. 12, p. 73-80.

Introduction to MPEG 7: Multimedia content description language. B.S. Manjunath, Philippe Salembier, Thomas Sikora (eds.). New York: Wiley, 2002.

Morse Edwin. [*et al.*]. *Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and "spring" displays.* «Journal of the American Society for Information Science and Technology», 53 (2002), n. 1, p. 28-40.

MultiMedia Information Retrieval: metodologie ed esperienze internazionali di content-based retrieval per l' informazione e la documentazione, [ed. da] Roberto Raieli – Perla Innocenti. Roma: AIDA, 2004.

Paquet Eric. [*et al.*]. *Virtualization, virtual environments and content-based retrieval of three-dimensional information for cultural applications*, in: *Proceedings of the SPIE, Videometrics VII, 21-23 January 2003, Santa Clara (CA).* Sabri F. El-Hakim, Armin Gruen, James S. Walton (eds). Bellingham (WA): SPIE, c2003, p. 137-147. (Proceedings of the SPIE, 5013).

Raieli Roberto. *MultiMedia information retrieval*, «Biblioteche oggi», 19 (2001), n. 10, p. 16-28.

Raieli Roberto. *Il sistema del Visual Retrieval*, «Bollettino AIB», 41 (2001), n. 1, p. 47-68.

Raieli Roberto – Grassi Antonio. *Principi avanzati di documentazione per le immagini biomediche*, «Giornale di gastroenterologia», 7 (2002), n. 3, p. 129-136.

Rolland Pierre-Yves. *Music information retrieval: a brief overview of current and forthcoming research*, in: *Proceedings of human supervision and control engineering of music*, Kassel, 2001, <www.engineeringandmusic.de/ individu/rollpier/ropiproc.html>

Smith John R. – Chang Shih-Fu. *Interoperable content-based access of multimedia in digital libraries*, in: *Proceedings of the First DELOS network of excellence workshop on information seeking, searching and querying in digital libraries*, Zurich, 2000, <www.ercim.org/publication/ws-proceedings/DelNoe01/>

Svenonius Elaine. *Access to Nonbook materials: the limits of subject indexing for visual and aural languages,*«Journal of the American Society for Information Science», 45 (1994), n. 8, p. 600-606.

Tzanetakis George – Cook Perry. *MARSYAS: a framework for audio analysis*, «Organised sound», 4 (2000), n. 3, p. 169-175.

The Viper Team – University of Geneva, *Viper's CBIRS page*, 2003, <viper.unige.ch/other_systems>

VIR Research group – University of Brighton, 2003, <www.cmis.brighton.ac.uk/Research/vir/VIR1.HTM>

The achievable innovation by the way of *MultiMedia Information Retrieval* (MMIR)

by Roberto Raieli and Perla Innocenti

Principles and practice of the MultiMedia Information Retrieval (MMIR), the organic complex of Visual Retrieval (VR), Video Retrieval (VDR), Audio Retrieval (AR) and Text Retrieval (TR) systems, are well known to computer scientists, engineers and mathematicians, but is now necessary that librarians too became familiar with MMIR technologies. The fields interested in the innovations of the MMIR are many and different, from the Medicine to the Geography, from the Engineering to the Visual Arts and the Music, each one introducing specific demands and challenges.

This articles outlines the state of the art of MMIR systems, with an introduction about the evolution of the multimedia retrieval concept across the classical Information Retrieval (IR) architectures of past and current multimedia archives. The point is that in the information searching area it may result limitative to operate in terms of a generic IR. In the traditional practice, every kind of documental search is compelled to the conditions of a search through textual language; it is necessary, on the contrary, to consider a broader criterion of MMIR, by which every kind of digital document is processed and searched through the elements of *language* more proper to its own nature. It is then possible to differentiate, in a more general methodology of multimedia searching, a method of TR based on textual information for the search of textual documents, from a method of VR based on visual data for the search of visual documents, of VDR based on video data for the search of videos, and of AR based on sounds for audio documents.

In databases where the content of the documents is substantially textual, it is appropriate that the keys of access will be terms and phrases extracted by the inside of that content; in multimedia databases, instead, it is inaccurate to attribute, from the outside, a textual description to contents that are well-grounded on a different structure of *sense*. Besides, if in the case of texts can be also suitable the method to analyse their *concept* and to attribute it a descriptor, this is not equally effective for images or sounds, where the subjective limits of the analysis are bigger, and not always concepts are of more interest than the concrete content of the documents, like forms, colours, movements, noises or music.

Such innovative and efficient systems have an information retrieval approach that treats directly the objective content of the documents, and for this it is defined

ROBERTO RAIELI, Biblioteca di Area delle Arti, Università degli Studi Roma Tre, e-mail raieli@uniroma3.it.

PERLA INNOCENTI, Centro Sistema Information Technology del Sistema Bibliotecario di Ateneo, Politecnico di Milano, e-mail perla.innocenti@polimi.it.

Bollettino **AIB**, ISSN 1121-1490, vol. 45 n. 1 (March 2005), p. 17-47.

content-based, in opposition to the traditional systems of indexing and searching based on terms describing of such material content, defined *term-based*. So, it is possible to retrieve multimedia documents by applying storing and retrieval techniques that operate directly on the audiovisual contents within database digital objects. Thanks to possibilities and tools offered by digital technologies, MMIR systems allow the retrieval of still images, audiovisual pieces and audio contents exploiting language-specific features of each document. According to similarity and other methods such as approximation and relationships of measures and values, users can perform queries by using figures, textures, shapes, colours, sounds, frames, movements, etc as retrieval keys.

The MMIR is a revolutionary *metasystem*, very specialized for an efficient treatment of digital multimedia objects. It is necessary to admit, however, that a good level of precision in the retrieval of documents can be reached only using in *combination* techniques and technologies of search based either on the definition of the concepts, through controlled terms, or on the representation of the content, through visual, audio and audiovisual elements. Both systems are able, in fact, to be harmonized. Term-based query can be a good preliminary method to select a part of the huge quantities of documents in regard to thematic areas, titles or authors. It can also be an ultimate way of cleaning up the inevitable specific *noise* of a content-based query. But, above all, the two procedures can operate in constant interaction, with an only search form, composing a query that by combining figures, sounds and texts is useful for searching very complex documents.

The article is divided in three parts dedicated to each MMIR sub-system, and introduce to Visual Retrieval, Video Retrieval and Audio Retrieval. Each part presents an overview on methods and techniques currently available and a brief survey of research areas and oputputs from national and international researchers, experts and scholars from the computer science and information retrieval community.

The scope of the authors is to contribute to the dissemination of MultiMedia Information Retrieval concepts and systems in Italy, envisioning their application by public and private organizations in a variety of fields within advanced documentation.