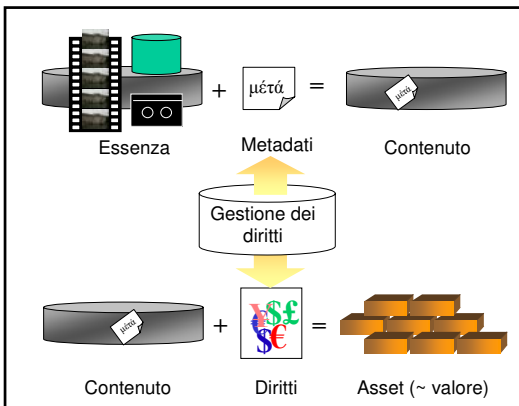
L' estrazione automatica dei metadati e loro affidabilita'

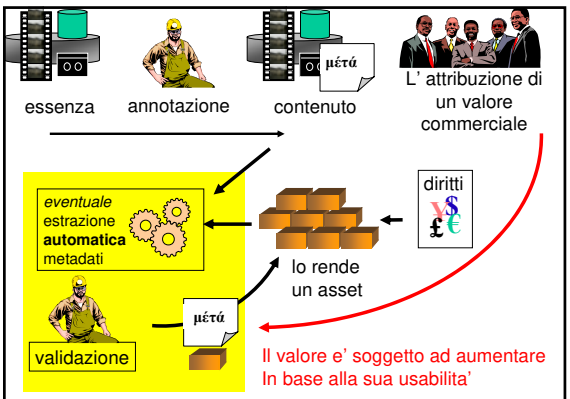
Igino Manfre'
Consulente in tecnologie televisive
Igino.manfre@gmail.com

μέτά

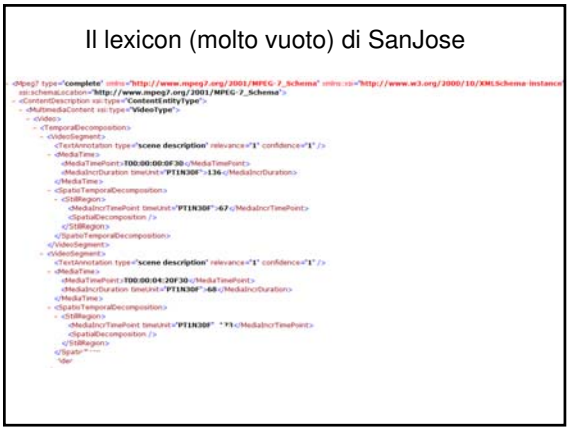
Oggi, ultimo incontro parliamo delle *finalita' del corso*

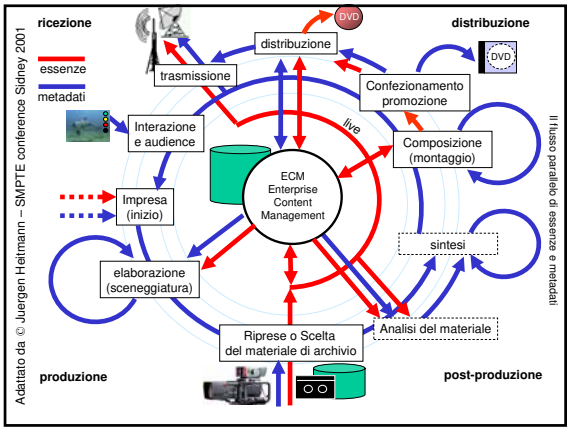
sopra



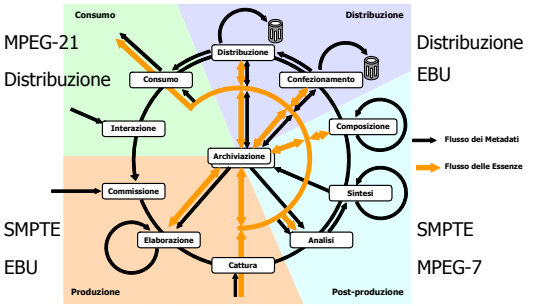


Il lexicon (molto vuoto) di SanJose





Quanto valgono i Metadati



AMWA

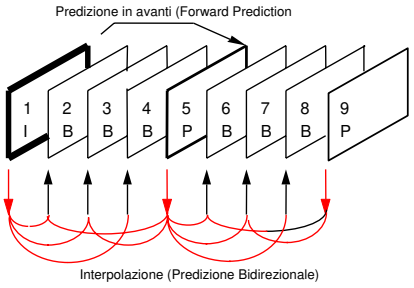
www.amwa.tv

Se tutte le imprese investono nella loro presenza in simili associazioni la ragione e' solo nei diritti

+

Arbitron, Inc.
Autodesk
Automatic Duck, Inc.
BPI Improve
BT Media & Broadcast
Digital Vision
E! Entertainment
eBus Limited
Marquis Broadcast
National Geographic
NBC Universal
Nielsen
Ninsight
OmniBus Systems
Panasonic
Perspective Media Group
Pro-Bel
SAIC
SGI Japan
VRT

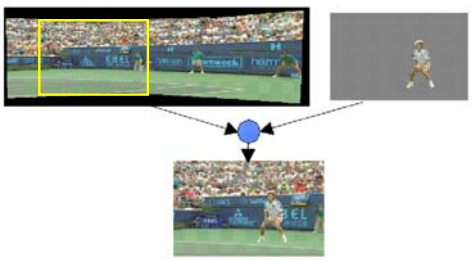
MPEG 1 – correlazione tra frame



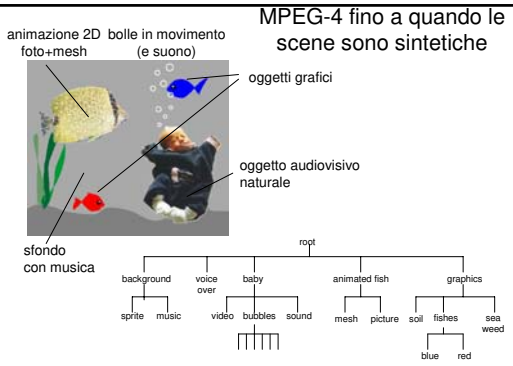
Per ogni blocco DCT 8x8 e' elaborato un vettore di moto che lo lega ai precedenti e successivi secondo la gerarchia del GoP, Group of Picture.

MPEG 4 – Analisi della scena

La compressione ad oggetti comporta il loro isolamento.



MPEG-4 fino a quando le scene sono sintetiche



© Rob Koenen 2000

Una volta specificate delle timeline e degli eventi …

Il problema e' la analisi della scena reali e la loro scomposizione in oggetti in moto

Ma nelle scene reali ?

per il video

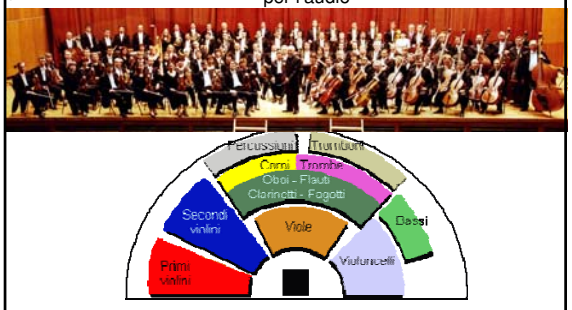
?

+

Il problema e' la analisi del panorama sonoro reale

Ma nelle scene reali ?

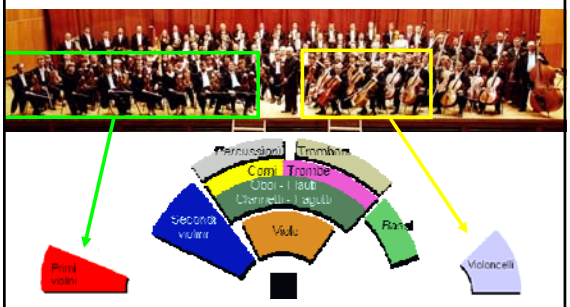
per l'audio



Il problema e' la analisi del panorama sonoro reale

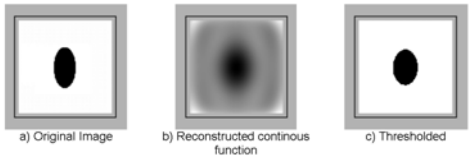
Ma nelle scene reali ?

per l'audio



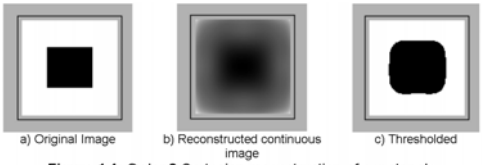
Come *estrarre* primi violini e violoncelli *distinguendoli* tra loro ?

se riusciamo ad ottenere …



a) Original Image b) Reconstructed continous function c) Thresholded

Figure 1.3: Order 8 Cartesian reconstruction of an ellipse.



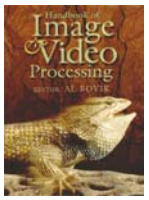
a) Original Image b) Reconstructed continuous image c) Thresholded

Figure 1.4: Order 8 Cartesian reconstruction of a rectangle

La estrazione automatica dei metadati

La estrazione automatica dei metadati non e' uno scherzo.
Quanto segue e' stato ampiamente prelevato da:

http://www.prestospace.org/project/public.en.html



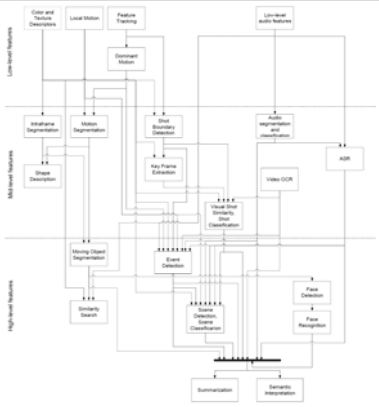
Handbook of Image and Video processing *edited by* Al Bovik

(Addison, 2000)

Scopo della analisi del contenuto

Lo scopo dell'analisi del contenuto e' quello di descrivere il materiale in modo compatto ed efficiente scambiabie e in modo trasferibile.

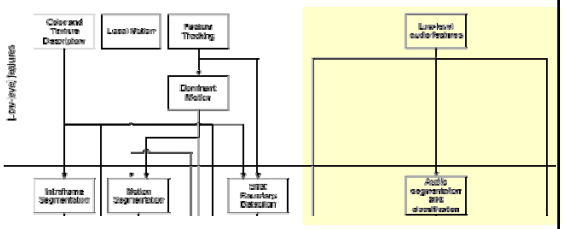
Il fine e' fare in modo che queste descrizioni non sono ambigue e che siano creabili senza altri dati oltre quelli del materiale stesso.



Correlazione tra i sistemi automatici di estrazione

Dal documento di PrestoSpace (D15-3, 2005) "**State of art of content analysis Tools for Video Audio and Speech**"

A basso livello (fisico)



Descrittori del colore e delle texture
Moto locale
Movimento del soggetto
Movimento complessivo (Global motion)
Caratteristiche audio a basso livello

Analisi a basso livello

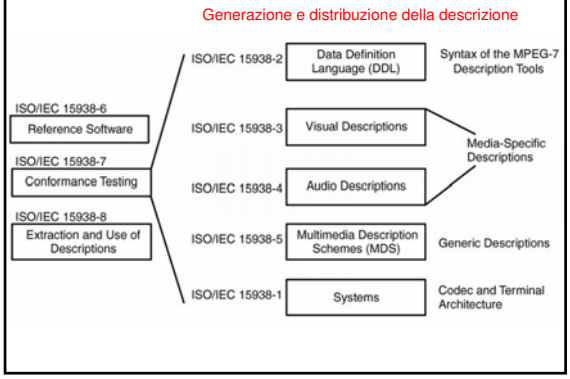
Le caratteristiche a basso livello sono singole proprieta' visuali legate al mezzo fisico come colore, texture, profilo, movimento. PrestoSpace li descrive come oggetto principale della nostro sistema visivo.

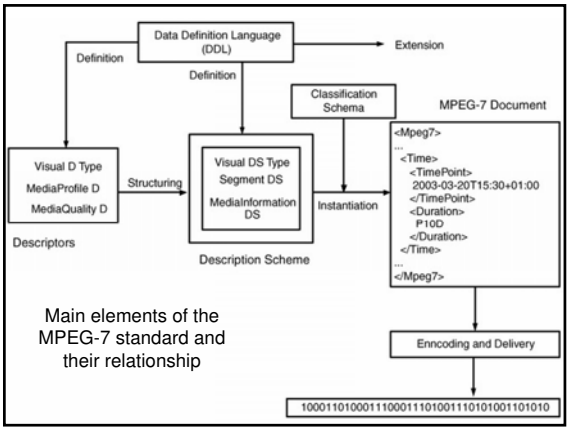
L'analisi a basso livello e' la base per un gran numero di algoritmi dei livelli superiori. Ogni descrittore consta di

- Rappresentazione: cosa e' contenuto nel descrittore e cosa significa.
- Estrazione (opzionale): come crearlo a partire da un dato visivo o sonoro
- Comparazione (opzionale): come determinare la collocazione nel relativo spazio, la distanza tra due descrittori dello stesso tipo.

L' MPEG-7 definisce un vasto set di descrittori visivi a basso livello.

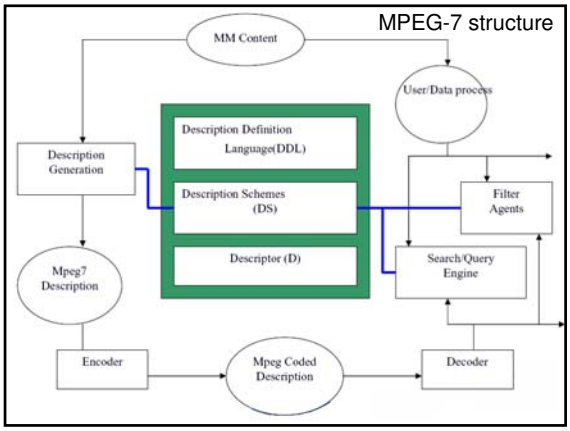
General Organization of MPEG-7





Main elements of the MPEG-7 standard and their relationship

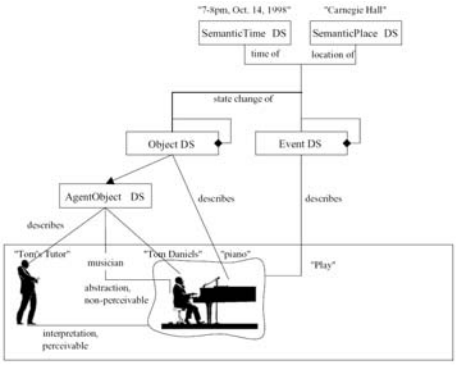
MPEG-7 structure



Moduli dell' MPEG-7

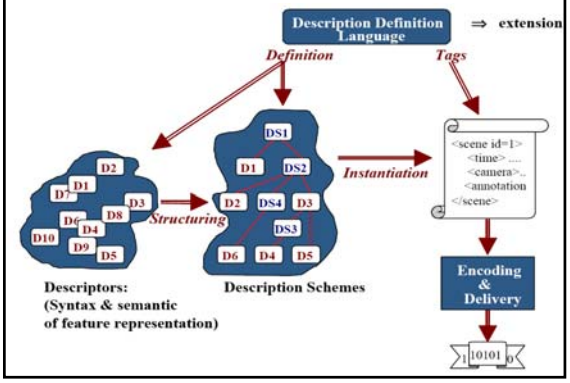
MPEG-7 modules	Functionality	Features
Descriptors (Ds)	Low level audio or visual features	color ,texture, motion, audio energy and so on.
Description schemes (DSs)	Higher level audio or visual features	region, segments, objects, events and immutable metadata
Description definition language (DDL))	Based on XML scheme language	coding schemes has both text and binary format

Esempio di descrizione concettuale con Semantic DS

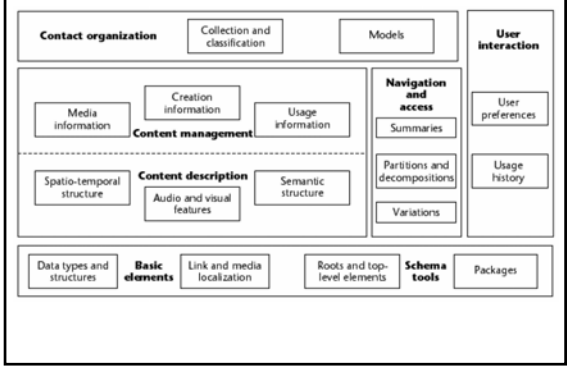


Narrative World

Relazione tra i componenti di MPEG-7

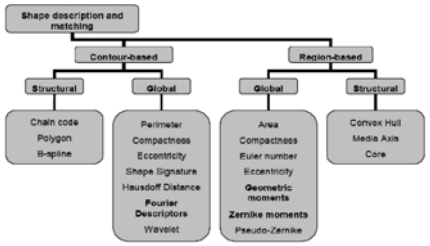


Overview dei Descriptor Tools



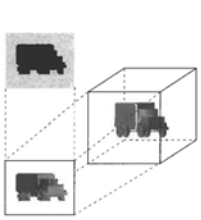
Contorno

L'MPEG4 ha introdotto la compressione ad oggetti, ma questa non e' stata applicata alle immagini naturali, preferendo in ogni caso un unico oggetto rettangolare in genere coincidente con il frame

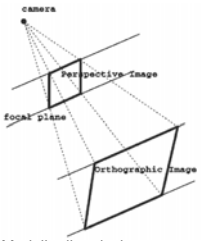


Tassonomia degli algoritmi per la rilevazione del contorno

Riconoscimento di forme



Modello di proiezione ortografica



Modello di proiezione prospettica

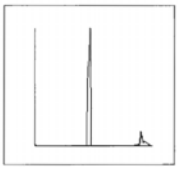
Immagine ortografica

Immagine prospettica

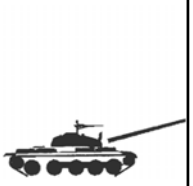
Riconoscere un oggetto

Simulazione di immagini ottenute da diversi tipi di sensori:

video

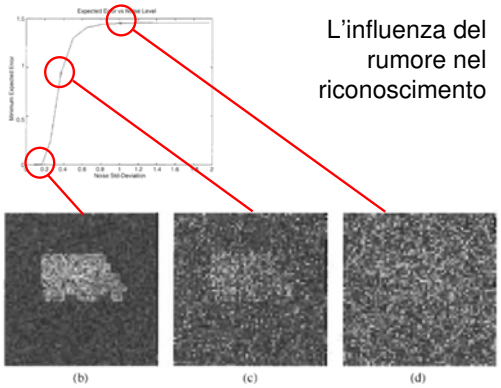


high range resolution range imagery (XPATCH, video aumentato)



Forward Looking InfraRed (FLIR)

L'influenza del rumore nel riconoscimento



A livello sonoro

E' analogo allo sforzo richiesto per comprendere una conversazione in presenza di un elevato rumore di fondo

Su un campo di sci

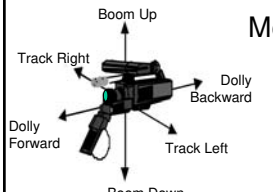
durante una manifestazione

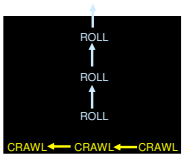
Movimento

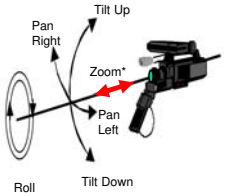
Si tratta di descrivere le caratteristiche di movimento degli oggetti nell'insieme della scena.
Il movimento possono essere relativo agli oggetti in primo piano (persone, edifici, etc) oppure al sistema di ripresa: cio' permette di descrivere lo sfondo in funzione dei movimenti di macchina (facilmente codificabili).
Tali relazioni possono essere estratte dalla analisi tridimensionale delle sequenze di immagini.
Vi sono sostanzialmente due tipi di moto:

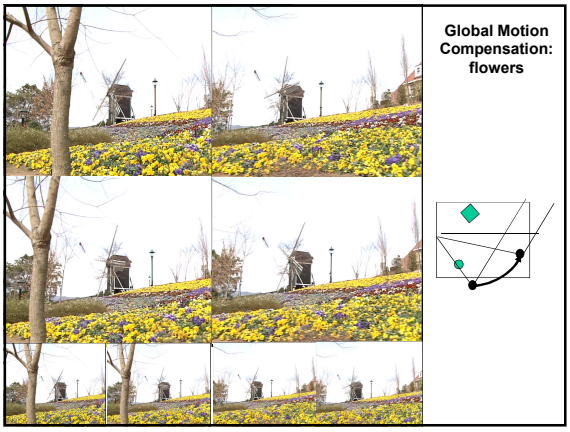
- **Movimento globale** (global motion, g.m.) il movimento di tutti i pixel di una immagine (caso tipico i titoli di coda)
- oppure il **moto relativo** di un oggetto rispetto ad altri o allo sfondo.

Movimenti di macchina

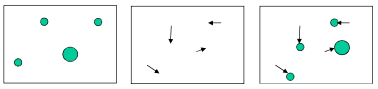




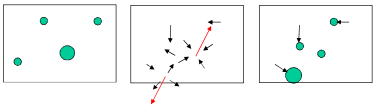




Modelli del moto



Modello translazionale (o non parametrico)



Modello parametrico
- affine (per oggetti lontani), 6 gradi di liberta'
- proiettivo lineare (in genere) 8

Stima e strategia di ricerca del moto

- Stima gerarchica su risoluzioni via via decrescenti (rende piu' facile l'analisi dei grandi movimenti, puo' essere applicato a qualsiasi modello)
- Flusso Ottico estensione dirette dei modelli di moto, idoneo per grossi movimenti orientati (appunto: flussi)
- Metodi basati sulla divisione in blocchi (tipici dell'MPEG1 e 2) che lavorano gerarchicamente alla ricerca di grossi movimenti di piccole aree.

- Nel dominio della frequenza la correlazione di fase (utilizzata anche nell'ultima generazione di codificatori MPEG2) e metodi di stima Bayesiani. Una implementazione a bassa risoluzione di questi produce il caratteristico ringing della compressione MPEG4

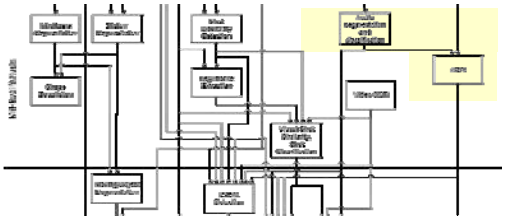
Analisi a basso livello audio

L'analisi a basso livello dell'audio e' molto sviluppata in quanto base per la i sistemi ad alto livello per il riconoscimento automatico del parlato (Automated Speech Recognition, ASR), funzione di analisi ad alto livello .
L'audio e' segmentato in piccole finestre (da 10 mSec in poi) ed e' poi processato sia nel dominio del tempo che in quello della frequenza o loro derivati.
Il dimensionamento di queste finestra puo' o meno rendere significativa l'elaborazione
Tra i domini piu' complessi c'e' il cepstrum, detto anche *spazio degli eventi*, sostanzialmente una trasformata di trasformata di fourier nel quale anche semplici operatori matematici (come il logaritmo) e' in grado di estrarre informazione paricolarmente significativa.

Analisi a basso livello audio (2)

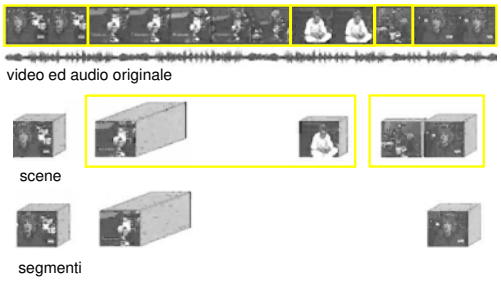
In particolare per quanto concerne gli stadi necessari alla conversione dei dialoghi (o del suono) al testo.
La discriminazione tra musica e parlato, tra strumenti e oratori diversi, tra strumenti diversi, anche della stessa famiglia.
La complessita' degli algoritmi spesso comunque impedisce – al momento - la elaborazione in tempo reale.
La comprensione del linguaggio umano e' una collezione di analisi interdisciplinare ad alto livello, e costituisce a tutti gli effetti uno dei sistemi piu' affidabili per la generazione automatica di metadati.
Sono disponibili sistemi ASR per tutte le lingue dei paesi avanzati piu' l'urdu (per scopi militari)

A medio livello - video



Segmentazione Intraframe e di moto
Rilevazione dell'inquadratura
Segmentazione e classificazione dell'audio
Descrizione dei contorni
Estrazione dei Key Frame
Video Optical Character Recognition (OCR)
Automated Speech Recognition (ASR)
Classificazione per similarita' di inquadratura

Analisi dell'audiovisivo



Qui sono rappresentate le scene come schede accodate
Come secondo passaggio si ricavano i segmenti.

Segmentazione spaziale, temporale e del moto

Analogamente alla creazioni di segmenti di scene riconosciute
A livello spaziale e' oportuno effettuare una segmentazione di tipo spaziale per evitare che dettagli di oggetti (ad esempio dei capelli bianchi, segmentazione cromatica) non seguano le traiettorie del soggetto cui appartengono.

Rilevazione della scena, storia

La detezione del cambio scena, tipica attivita' a basso livello, e' opportuno sia seguita da segmentazione (livello medio) allo scopo di ricomporre sequenze di cambi scena coerenti, una attivita' moderatamente inversa.
Caso tipi di errore dati dalla errata segmentazione e' lo scatto di un flash (per non citare una lampada stroboscopica) che nelle prime compressioni MPEG2 facilmente produce effetti devastanti.
Possiamo collocare a livello medio la rilevazione della scena, mentre a livello alto la estrazione della storia peraltro con variabilita' estremamente elevata, spesso ambigua.
Parafrasando Eijsenstein, lo stesso girato montato in modo diverso e' un film diverso.

Riconoscimento di testi in scena

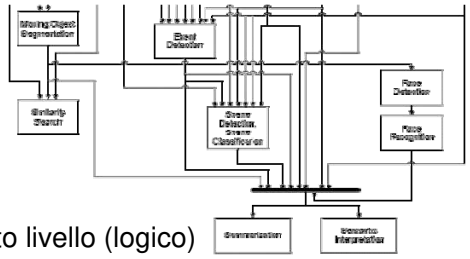
Simile ma piu' complesso del normale OCR (Optical Character Recognition) per la piu' bassa risoluzione e frequente interlacciamento, senza contare il posizionamento nello spazio della scritta, che in un ordinario OCR porta spesso al fallimento del riconoscimento.

Funzioni audio a livello intermedio

Una categorizzazione dei trattamenti sonori a livello intermedio possono consistere in metodi per la:

- segmentazione e classificazione (comp. semplice)
- clusterizzazione (raggruppamento, comp semplice)
- riconoscimento di campioni (pattern, ad esempio riconoscimento di un oratore tra molti, dipendente dalla dimensione del database)

Questo ovviamente tanto nel dominio del tempo che in quello frequenza ma facilmente estendibile o in qualsiasi altro derivato.



Ad alto livello (logico)

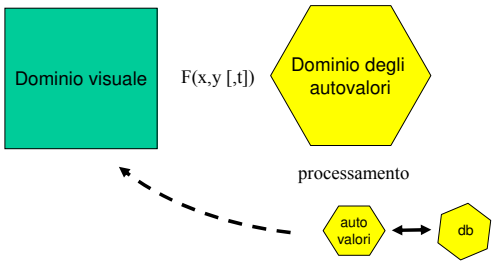
Segmentazione di oggetti in movimento
Ricerca per similarita' (distanza in opportuni spazi derivati)
Rilevazione di eventi, scene e loro classificazione
Rilevazione e riconoscimento facciale
Sinossi – Riassunto (Interpretazione Semantica)

Il riconoscimento del volto umano

Lo sviluppo di un modello computazionale per il riconoscimento degli oggetti naturali come il volto umano e' abbastanza difficile perche' il nostro riconoscimento e' basato su complessi stimoli significativi multidimensionali
Piu' che altro si puo' lavorare sulla densita' di probabilita' di apparizione (Probability Density Function, PDF) di un volto umano in due dimensioni.

Immagine da analizzare | Rilevazione del volto | Centratura del volto | Rilevazione della espressione facciale

Una trasformazione



alla fine e' sempre una compressione

si tratta di ridurre l'informazione alla sola significativa

Immagine originale allineata | Immagine ricomposta dagli autovalori (**85 byte**) | Immagine compressa in JPEG (530 byte)

Caratteri intrapersonali e extrapersonali

Gli autovalori possono far parte di due classi diverse

autovalori **intrapersonali** se le loro differenze' possono essere causate da sottili variazioni di espressione o illuminazione della ripresa

autovalori **extrapersonali** se tali diversita' fanno riferimento a differenze sostanziali (come il calvizie, presenza o meno di barba o occhiali etc).

Non sempre un volto intero

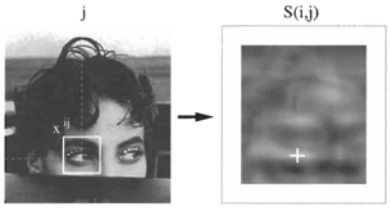
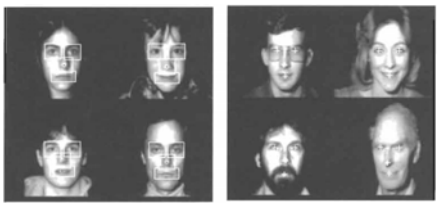


Immagine in ingresso e a destra la probabilita' S(i,j) che il rettangolo comprenda un occhio sinistro

E' possibile **limitare l'esame a piccole aree della immagine** (gli occhi, la bocca, posizioni mutue con la punta del naso) che producono degli autovalori particolari, detti *eigenfeature* legati alle **espressioni**

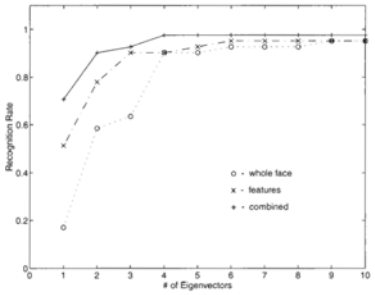
Aspetto ed espressione

macchine da addestrare



A sinistra addestramento al riconoscimento delle **espressioni** effettuato limitando l'esame a piccole porzioni della immagine e a destra risultato del riconoscimento attraverso tale addestramento

Piu' informazione, migliori risultati



Aree per il riconoscimento delle espressioni e rapporto di corretto Riconoscimento migliore in caso di uso di pochi autovalori misti

autovalori diversi per volti e espressioni

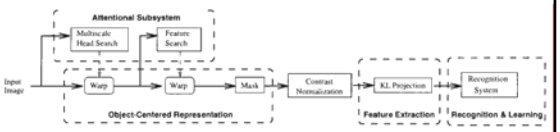
Immagini in test

Volti riconosciuti

Espressioni riconosciute

Workflow

Ricerca della testa in diverse inquadrature (multiscale)



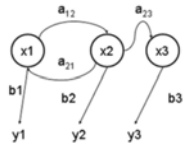
92% di risultati positivi

Alcune delle foto utilizzate per testare l'accuratezza del riconoscimento facciale.
Il successo e' stato del 92%, l'errore del riconoscimento ha presentato una deviazione standars di 15 gradi

Trascrizione dei dialoghi

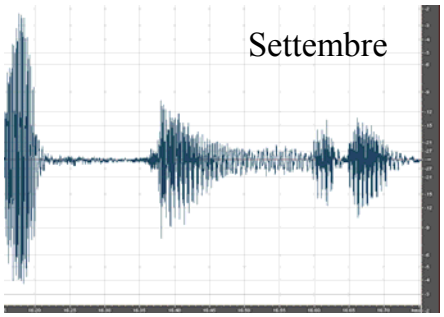
Le applicazioni di riconoscimento automatico del parlato (ASR), hanno goduto di notevoli finanziamenti di provenienza militare, e solo negli ultimi 10 anni dalla unione europea (dove si parlano 15 lingue!).
Per le applicazioni relative alla metadatazione non possono essere utilizzati gli algoritmi che fanno uso di raffronto con un vocabolario di fonemi, e neanche sistemi che necessitino di addestramento.
I sistemi in uso si basano sui modelli markoviani nascosti (Hidden Markov Model, HMM), dei quali il riconoscimento vocale e' stato negli anni '70 una delle prime applicazioni.

Andrey Markov (1856 – 1922)

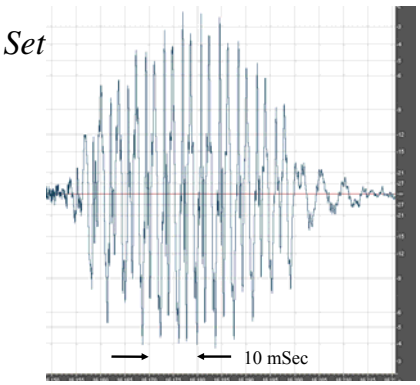


Un **processo di Markov** è un processo stocastico nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende **unicamente dallo stato di sistema immediatamente precedente** e *non* dal come si è giunti a tale stato
Una **catena di Markov** è un processo di Markov a stati discreti, ovvero è un processo stocastico discreto per cui ad ogni istante t estraiamo dal processo una variabile casuale discreta
Un **Modello Nascosto di Markov** è una catena di Markov i cui stati non sono osservabili direttamente.

Una delle possibili ragioni dell'uso dei modelli nascosti di Markov e' che il discorso umano puo' essere considerato come una sequenza eventi stazionari di breve durata, ad es. 10 millisecondi.

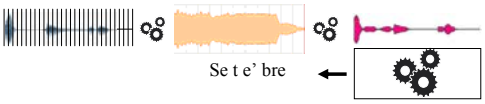


Milena Gabanelli che pronuncia *settembre* in 600 msecondi



Si, ma come funziona?

Semplificando, si utilizza una FFT (Fast Fourier Transform) con una finestra da 10 mSec, e si produce un vettore di coefficienti cepstrali (correlazione DCT della trasformata di Fourier), L'applicazione di un modello nascosto di Markov elaborando i piu' significativi di questi valori (i primi, quelli piu' lontani dalla specificita' del singolo speaker) produce per ogni fonema una distribuzione di probabilita' dipendente (al piu') a fronte del precedente. Concatenando questa attivita' si puo' contare di riconoscere il discorso.



C'e' da considerare che poiche' il riconoscimento automatico del parlato ha una **grande importanza nel settore militare**, si e' raggiunta l'eccellenza in quasi tutte le lingue del primo mondo… … ed in molte del terzo…

Non perdere mai di vista l'obiettivo

3.1 Basic definitions

Let $\mathbb{A} \subset \mathbb{N}$, $\mathbb{A} = 0, \ldots, H-1 \times 0, \ldots, W-1$, where H and W are the height and width of a frame, respectively.

Definition 3.1 (Frame) *A* frame f_t *is a function from* $\mathbb{A}$ *to* $\mathbb{N}$ *where for each spatial position* (x,y) *in* $\mathbb{A}$, $f_t(x,y)$ *represents the value of the pixel* (x,y) *at time* t.

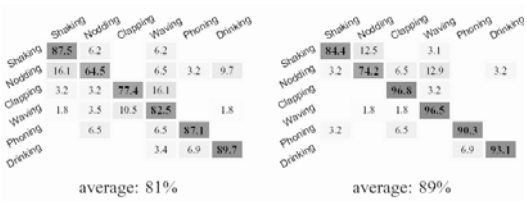
Definition 3.2 (Video) *A* video V *is a sequence of frames* f_t *with an accompanying audio track and can be described by* $V = (f_t)_{t \in [0, T-1]}$, *where* T *is the number of frames contained in the video. The number of frames is directly associated with the frequency and the duration of visualization.*

Il formalismo matematico e' una necessita': qui non si tratta di contestarlo, ma per favore, una volta partorite definizioni come queste qua sopra, tornate alla realta', pensate alla pellicola, al fotogramma, alla scansione, alla cromaticita'.

Database degli atteggiamenti

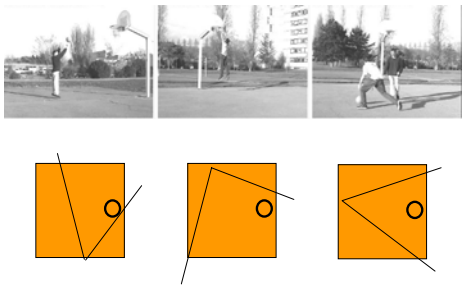
Esempio di ciascuna classe del database degli atteggiaenti elaborato dal progetto europeo MUSCLE (**M**ultimedia **U**nderstanding through **S**emantics, **C**omputation and **Le**arning, progetto IST FP6-507752, da "Progress on Applications of Machine Learning Techniques" (*Deliverable:* D8.1c, 2005)

…e sua efficacia



Matrice di confusione per il database video degli atteggiamenti pesati uniformemente (a sinistra) e in base alla classe (a destra).
E' difficile sapere cosa e' vero in queste dichiarazioni e cosa e' artefatto.
Concediamoci il diritto di poter verificare in pratica qualora servissero.

Inquadratura



Esempi di ciascuna delle tre classi del database "Basketball" di MUSCLE

Fine

Spero abbiate potuto trovare degli spunti interessanti.

Ricordate che la curiosita' e' la molla della invenzione.

Mandero' una e-mail a tutti per informarvi quando sara' possibile scaricare il materiale
(stato in cui rimarra' per un paio di settimane).

Igino.manfre@tiscali.it