Escola Politècnica Superior de Castelldefels
UNIVERSITAT POLITÈCNICA DE CATALUNYA

# PROJECTE DE FI DE CARRERA

**TÍTOL DEL PFC: Anàlisi audiovisual mitjançant descriptors MPEG-7**

**TITULACIÓ: Enginyeria de Telecomunicació (segon cicle)**

**AUTOR: Enric Bosch Anglarill**

**DIRECTOR: Francesc Tarrés Ruiz**

**DATA: 24 febrer de 2006**

**Títol:** Anàlisi audiovisual mitjançant descriptors MPEG-7

**Autor:** Enric Bosch Anglarill

**Director:** Francesc Tarrés Ruiz

**Data:** 24 de febrer de 2006

**Resum**

L'objectiu és dur a terme una classificació audiovisual de dibuixos animats a través de descriptors MPEG-7.

Inicialment s'ha creat una base de dades audiovisual sobre dibuixos animats formada per 140 fitxers (de vídeo, d'àudio i de text).

A continuació s'ha implementat el descriptor de textura HTD definit per l'estàndard MPEG-7. Aquest descriptor analitza la potència de la imatge en direccions i bandes en el domini freqüencial.

S'han extret de cada segment de vídeo les característiques associades al descriptor de textura.

Aquestes característiques s'han utilitzat per introduir-les a quatre classificadors (PCA, LDA, M-NN i Fuzzy Integral) i obtenir els resultats de la classificació de vídeos de dibuixos animats.

Per tal de dur a terme una classificació multimodal, s'ha adjuntat el descriptor de textura amb altres descriptors (àudio, moviment i color) .

S'han generat fitxers de classificació amb 43 característiques d'àudio (volum mig, creuaments per zero, ample de banda...) i 79 de vídeo (mitja RGB, autocorrelograma, filtre Gabor nº1, ....).

A partir d'aquests fitxers s'han fet diverses proves de classificació.

En el cas d'utilitzar només el descriptor de textura, els resultats han estat d'un 77% de reconeixement de vídeos de dibuixos animats.

Utilitzant característiques d'àudio i vídeo, el resultat de reconeixement de dibuixos en el millor cas és del 85 %.

**Title:** Audiovisual analysis through MPEG-7 descriptors.

**Author:** Enric Bosch Anglarill

**Director:** Francesc Tarrés Ruiz

**Date:** February, 24th 2006

**Overview**

The objective of this master thesis is to perform a cartoon audiovisual classification using MPEG-7 descriptors.

Firstly, an audiovisual database about cartoons and other television genres has been created. This database consists of 140 segmented files (video, audio and text) which correspond to more than 2 houres of video.

Cartoons are supposed to have flat and homogeneous textures. Thus, in the second step of this work the Homogeneous Texture Descriptor (HTD) defined in the MPEG-7 standard has been implemented. This descriptor analyzes the energy image in spatial directions and frequency bands.

The texture features included in this descriptor has been computed for the whole audiovisual database and utilized as input for four classifiers (PCA,LDA,M-NN and Fuzzy Integral) with the objective of cartoons classification.

A first classification has been carried out considering only the texture features. After analyzing these preliminary results, the texture descriptor has been included in a multimodal framework (texture,audio, movement and color).

Results have shown tha the texture descriptor alone obtains a high classification rate (above the 77%) but this can be improved till a detection rate of 85% when adding some audio and color video features.

Aquest projecte està dedicat:

A la família i a la Sara,
pel suport moral.

Al Toni i Francesc,
per les seves inspiracions.

# ÍNDEX


**INTRODUCCIÓ** ..... **1**

**CAPÍTOL 1. SITUACIÓ INICIAL** ..... **6**

**1.1. Contingut audiovisual** ..... **6**

**1.2. Visió global** ..... **7**

**CAPÍTOL 2. ANÀLISI MULTIMODAL** ..... **8**

**2.1. Estàndard MPEG-7** ..... **9**

**2.2. Àudio** ..... **11**
2.2.1. Característiques a nivell de trama ..... 11
2.2.2. Característiques a nivell de clip ..... 13

**2.3. Vídeo** ..... **13**
2.3.1. Color ..... 13
2.3.2. Textura ..... 15
2.3.3. Forma ..... 15
2.3.4. Moviment ..... 15

**CAPÍTOL 3. FITXERS DE CARACTERÍSTIQUES** ..... **17**

**3.1. Segmentació manual** ..... **17**

**3.2. Etiquetatge manual** ..... **18**

**3.3. Exportació dels fitxers** ..... **18**

**3.4. Extracció de característiques** ..... **19**

**3.5. Generació dels fitxers d'entrenament** ..... **19**

**CAPÍTOL 4. CARACTERÍSTICA DE TEXTURA** ..... **20**

**4.1. Homogeneous Texture Descriptor** ..... **20**
4.1.1. Sistema de visió humana (HVS) ..... 20
4.1.2. Descripció del descriptor ..... 21

**CAPÍTOL 5. CLASSIFICADORS** ..... **26**

**5.1. PCA** ..... **26**

**5.2. LDA** ..... **29**

**5.3. M-NN** ..... **30**

**5.4. Fuzzy Integral** ..... **31**

**CAPÍTOL 6. PROGRAMES ..... 32**

**6.1. MacroFitxer ..... 32**
6.1.1. Funcionament ..... 32

**6.2. Descriptor de Textura ..... 33**
6.1.1. Eines utilitzades ..... 33
6.1.2. Estructura ..... 34
6.1.3. Funcionament ..... 39

**CAPÍTOL 7. RESULTATS ..... 41**

**7.1. Segments de vídeos ..... 41**

**7.2. Extraccions de característiques ..... 41**

**7.3. Classificació de dibuixos animats ..... 42**
7.3.1. Classificació amb el descriptor de textura ..... 43
7.3.2. Classificació amb el descriptor de textura reduït ..... 44
7.3.3. Classificació multimodal ..... 45
7.3.4. Classificació multimodal amb finestra temporal ..... 46

**CAPÍTOL 8. CONCLUSIONS I LÍNIES FUTURES ..... 48**

**BIBLIOGRAFIA ..... 50**

**ANNEXOS ..... 51**

**1.1. Descripció de l'aplicació d'extracció de característiques ..... 51**

**1.2. Proves de verificació del funcionament de la FFT ..... 57**

**1.3. Fitxer de característiques del descriptor de textura reduït ..... 59**

# INTRODUCCIÓ

En l'actualitat es disposa de gran volum d'informació, però un dels problemes és cercar la que interessa d'una manera ràpida i eficaç.

Com a conseqüència del desenvolupament tecnològic i de la televisió digital, la informació tendeix a ser multimèdia (en detriment de la informació textual), sent més directa i vistosa. En la Taula. 0.1 s'observa el consum anual de bits als Estats Units que és d'aproximadament 230 Exaytes cada any.

**Taula. 0.1** Consum anual de bits als Estat Units

| **Un total de 70 M de cases** | **≈ 230 Exabyte / any** |
|---|---|
| ***Televisió*** | ***94%*** |
| *Ràdio* | *1,7%* |
| *Música Gravada* | *0,4%* |
| *Diaris* | *0,0003%* |
| *Llibres* | *0,0002%* |
| *Revistes* | *0,0002%* |
| ***Vídeo a casa*** | ***3,3%*** |
| ***Jocs de vídeo*** | ***0,6%*** |
| *Internet* | *0,0003%* |
| [Font: UC Berkley: How much information] | |

Analitzant aquestes dues tendències ha sorgit la necessitat de buscar sistemes de gestió i classificació en continguts multimèdia (processament i indexació automàtica). És a dir, disposar de continguts audiovisuals convenientment catalogats de manera que es puguin personalitzar d'acord amb el perfil de l'usuari (notícies, esports, dibuixos animats etc.), dur a terme una recerca de continguts a partir de paraules clau (casa, cotxe, arbre, camp etc.) etc.

Per a la classificació dels continguts multimèdia de manera automàtica és necessària l'anàlisi de tot tipus d'informació: imatges, sons, texts, paraules... En principi, un sistema d'anàlisi de contingut multimèdia serà més eficient si l'anàlisi és multimodal, (si s'analitzen característiques d'àudio i vídeo) ja que s'espera que disminueixi la probabilitat d'error en la seva classificació. Els sistemes d'anàlisis multimodals són relativament nous, ja que per analitzar senyals d'àudio i de vídeo simultàniament es requereix molta capacitat computacional, i també per la poca cooperació que tradicionalment existia entre les comunitats científiques dels dos camps (vídeo i àudio).

A partir de la necessitat de descriure completament els continguts i/o materials audiovisuals, va néixer l'estàndard MPEG-7. El MPEG-7 estandarditza la forma en què el contingut pot descriure's, és a dir, estableix una sintaxi estàndard de manera que sigui possible la interoperatibilitat entre diversos sistemes de catalogació o recerca de continguts. Per tant, és interessant conèixer aquest

estàndard i tenir-lo com a referència en l'elaboració d'aquest projecte (veure [1]).

La descripció de contingut audiovisual pot ser a diferents nivells i es tracta d'una estructura jeràrquica. Es poden extreure característiques de baix nivell semàntic representatives del segment (color dominant, tipus de moviment, formes, textures etc.), característiques de mig nivell d'acord amb les de baix nivell (classificació de notícies, pel·lícules, esports, etc.) i característiques d'alt nivell com per exemple "partit de futbol Barça-Madrid" o "Gol d'Eto rematant de cap" que podrien agrupar i interpretar les característiques dels dos nivells anteriors.

Aquest projecte està englobat dins de l'esquema presentat en la Fig.0.1 que té com a objectiu fer una classificació de gèneres de TV (nivell semàntic mig) utilitzant característiques o descriptors d' àudio i vídeo. En la part d'àudio, les característiques es poden diferenciar entre característiques a nivell de trama i característiques a nivell de clip (2.2). En la part de vídeo (2.3), es fa la divisió en característiques de moviment, color, textura i forma que s'explicaran amb més detall en apartats següents.

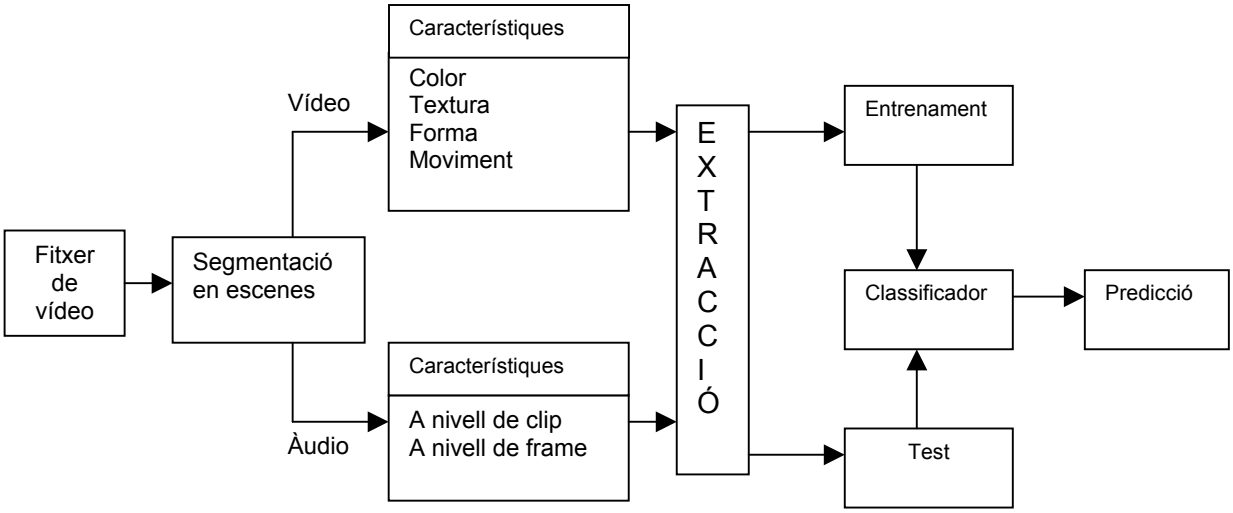



**Fig.0.1** Esquem de l'estructura d'un sistema de classificació

Una vegada s'ha fet la segmentació del vídeo en escenes, es procedeix a l'extracció d'aquestes característiques. Les característiques extretes en tots els camps (àudio i vídeo) es guarden en uns fitxers de text per ser utilitzats posteriorment.

Aquestes característiques s'utilitzen per entrenar un classificador i per fer un testeig una vegada entrenat. Per tant, el classificador és l'encarregat de dur a terme la classificació del contingut multimèdia i dir a quin gènere correspon aquest contingut (notícies, dibuixos, pel·lícula, esports etc.).

Aquest projecte es centrarà en el gènere de dibuixos animats. S'ha escollit aquest gènere per les múltiples aplicacions pels nens petits, com ara la

gravació automàtica d'aquest gènere, habilitar que només es puguin veure dibuixos animats per evitar les escenes de violència i sexe, seleccionar automàticament el canal on facin dibuixos animats etc. És una manera fàcil de filtrar la programació que volem que vegin els nostres nens per a una millor educació.

Per l'anàlisi d'aquest gènere s'ha implementat un descriptor de textura ja que es pensa que és una característica important per classificar els dibuixos animats de manera eficient. El descriptor de textura és el HTD (Homogeneous Texture Descriptor) que es basa en filtres de Gabor. Aquests calculen l'energia mitja d'una sèrie de canals de dues variables (direcció i banda de freqüència) de la FFT 2D d'una imatge.

Per dur a terme la classificació de dibuixos animats a través del descriptor de textura HTD s'han seguit els passos de la Fig.0.2 .

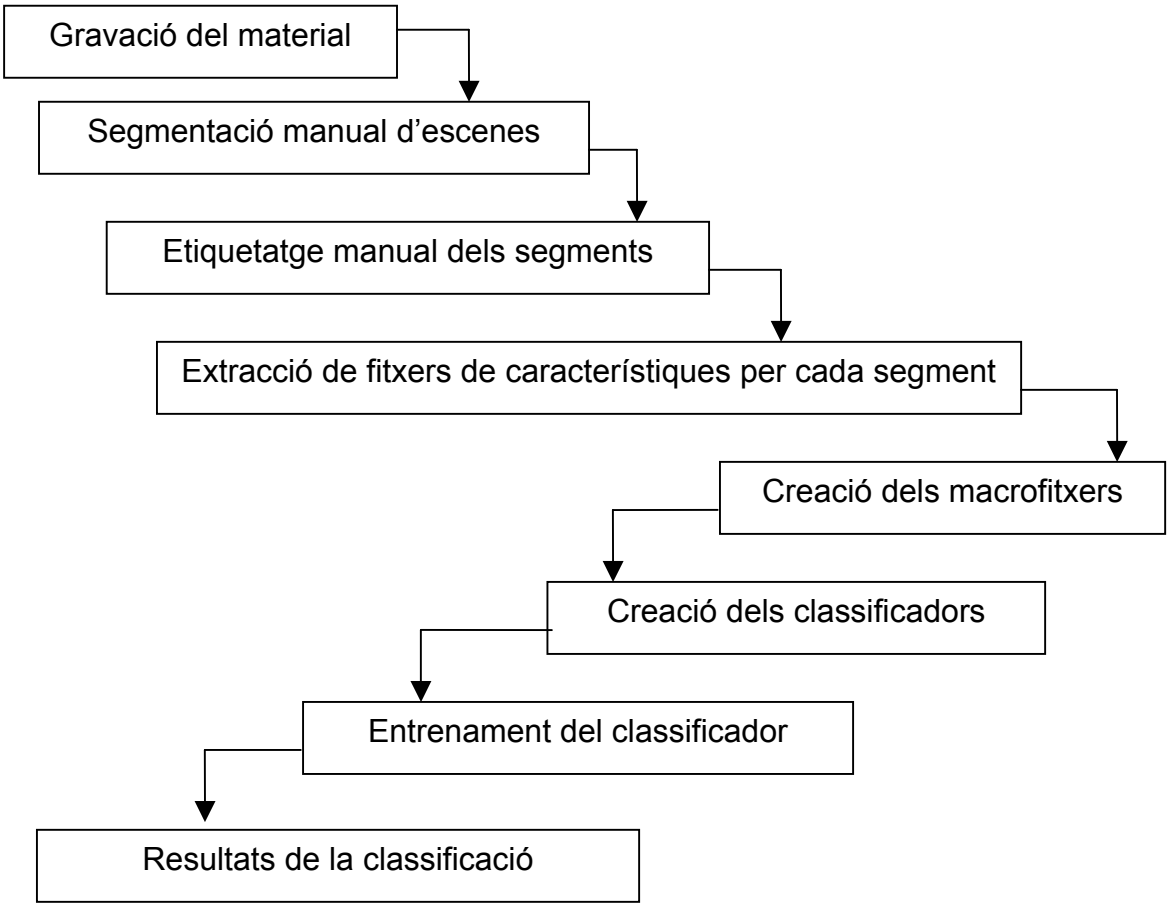



**Fig.0.2** Tasques seguides fins a la classificació

Per seguir aquests passos ha estat necessari implementar unes eines per la creació de macrofitxers, i implementar uns classificadors en Matlab.

Per a realització d'aquest projecte s'han utilitzat diverses eines i llibreries que han facilitat la tasca. L'eina utilitzada per segmentar els vídeos de dibuixos animats en escenes és el Adobe Premier 6.5, el llenguatge de programació utilitzat el C++, l'entorn de programació el Visual Studio .NET 2003, les llibreries utilitzades el OpenCV, IPL i IPP de Intel que ofereix funcions de processat d'imatge i vídeo.

A continuació comentarem breument el contingut de cada capítol.

En el primer capítol s'explica perquè és necessària una classificació de continguts audiovisuals, i la situació inicial en començar el projecte.

En el segon capítol s'explica què és l'anàlisi multimodal, la seva estructura, i les seves característiques.

En el tercer capítol es detallen els passos per obtenir un fitxer de característiques del vídeo a analitzar, i el seu posterior entrenament.

El quart capítol conté una explicació del descriptor Homogeneous Texture Descriptor que és la base del projecte, mentre que en el cinquè capítol es defineixen els quatre classificadors que s'han implementat i utilitzat per la classificació del gènere de dibuixos animats.

En el sisè capítol es comenten els programes realitzats, concretament la classe CVideoTexture, CGaborFilter, l'entorn gràfic i altres petites aplicacions realitzades.

En el setè i vuitè es presenten els resultats i conclusions.

# CAPÍTOL 1. SITUACIÓ INICIAL

Heu pensat mai com obtenen els programes de televisió una llista de tots els documentals que parlen sobre lleons? O com cerquen una llista de cançons a partir de l'extracte de la veu de l'Elvis Presley? O com es pot gravar una pel·lícula de la televisió filtrant els anuncis? En aquest apartat es comentaran les eines per poder contestar aquestes preguntes. També s'explicarà l'estat inicial del projecte i el seu objectiu.

## 1.1. Contingut audiovisual

La informació audiovisual (documentals, imatges, sons etc.) constitueix una font de riquesa i coneixement. La literatura s'ha assegurat la conservació en museus i biblioteques, però la producció audiovisual encara no es considera tan important.

El procés de digitalització que s'està portant a terme en la societat, especialment la digitalització per la conservació de la informació, fan que es busquin sistemes automàtics i d'indexació de la informació audiovisual en bases de dades, ja sigui per ús professional, educació i investigació.

En principi, els estudis sobre el vídeo digital es van centrar sobre la compressió de la informació (tasca assumida pel Motion Picture Expert Group (MPEG) i Joint Picture Expert Group (JPEG)), creant estàndards reconeguts com el MPEG-1,2 i el JPEG i JPEG2000. Però ara el punt important és la cerca d'informació audiovisual. El MPEG-7 va donar la solució a aquesta necessitat, on el seu objectiu és estandarditzar els descriptors audiovisuals permetent una cerca eficient de la informació multimèdia.

És important remarcar que l'estàndard MPEG-7 està enfocat a la indexació del contingut audiovisual, i deixa més oberta la part de classificació. És a dir, defineix una sèrie de descriptors, però no defineix cap tipus de classificador, ni especifica la importància de cada descriptor pels diferents tipus de contingut audiovisual.

En la següent imatge s'observen els objectius que busca MPEG-7. El primer pas consistiria en una anàlisi del document multimèdia i les relacions entre els elements. Aquesta anàlisi es podria fer a mà o mitjançant una eina informàtica. Per això MPEG-7 defineix una sèrie de descriptors estàndards, que poden ser ampliats. Alguns d'ells són: estructures bàsiques, descriptors de color, descriptors de textura, descriptors de forma, descriptors de so, etc. També es poden utilitzar eines de anotació per descriure la semàntica del document.

D'aquesta manera es pot ser capaç de desenvolupar potents buscadors o classificadors de documents multimèdia.

Aquesta estructura serà similar a la utilitzada en aquest projecte on s'extreuen característiques, s'utilitzen descriptors MPEG-7 i es fa una classificació.

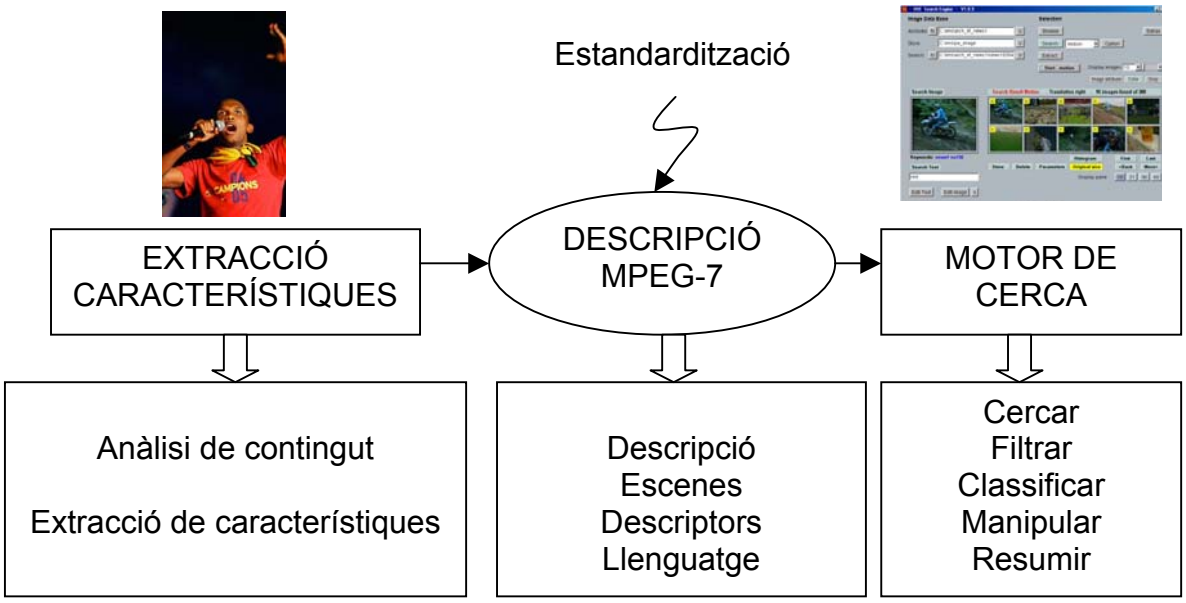



**Fig. 1.1.1** Objectius de MPEG-7

## 1.2. Visió global

Aquest PFC forma part d'un projecte més global que té com a objectiu extreure característiques de contingut audiovisual analitzant diferents nivells (veure Fig.0.1 ) i fer la classificació de gènere.

Aquest projecte global analitza característiques d'àudio i vídeo, detallades més endavant. S'ha disposat d'aquestes característiques implementades per realitzar el PFC.

Algunes d'aquestes característiques estan definides en l'estàndard MPEG-7 i d'altres, per la seva simplicitat i validesa, també es va creure oportú implementar-les en el seu moment encara que no formessin part de MPEG-7.

En els apartats 2.2 i 2.3 es detallen algunes d'aquestes característiques de manera més àmplia.

L'objectiu principal d'aquest PFC ha sigut analitzar e implementar un nou descriptor de textura i incloure-ho dins d'un context multimodal amb l'objectiu de classificar dibuixos animats. S'ha escollit aquest gènere per les múltiples aplicacions (filtració de continguts violents i sexuals, gravació automàtica de dibuixos animats, personalització de preferències etc.) .

S'ha implementat una característica de textura perquè es creu que és important per discriminar entre dibuixos animats i altres gèneres de TV. Per tal de fer la classificació del gènere de dibuixos animats s'han seguit vuit passos, detallats a la Fig.0.2 i comentats amb més detall al capítol 3.

# CAPÍTOL 2. ANÀLISI MULTIMODAL

Com a conseqüència dels avanços científics en l'emmagatzematge i les xarxes ens dirigim cap un món amb abundant contingut multimèdia. Per tant, necessitem indexar el contingut per a la seva cerca, manipulació i classificació.

Un mètode per la classificació d'un document audiovisual és segmentar la part de vídeo, segmentar la part d'àudio i després integrar-ho. En el projecte s'ha utilitzat l'estructura de la Fig. 2.1 .

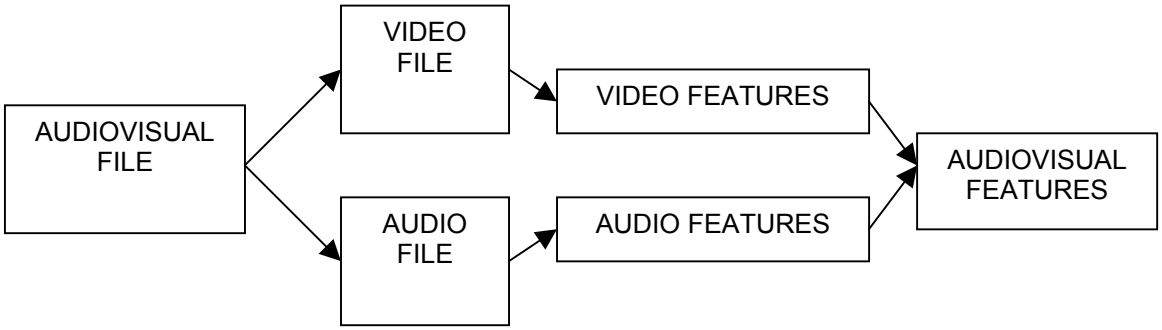



**Fig. 2.1** Anàlisi multimodal

Un dels pioners en la classificació multimodal de contingut audiovisual va ser Fisher (veure [3] ) que va classificar programes de televisió en notícies, curses de cotxes, tennis, anuncis i dibuixos animats utilitzant tres nivells d'extracció. En el primer nivell es van extreure algunes propietats sintàctiques del vídeo (l'histograma de color, el moviment, la forma de veu i l'espectre de l'àudio). En el segon nivell, les característiques d'àudio i de vídeo es van utilitzar per determinar l'estil de cada escena. I l'últim nivell, la variació temporal de cada característica es compara amb uns fitxers de característiques predefinits que corresponen a cada gènere de programa.

Wang també va implementar un classificador de gèneres de programa utilitzant HMM (Hiden Markov Models) (veure [4] ). Els autors expliquen la seva motivació per la utilització HMM pel fet que en tot moment els valors de les característiques de diferents programes poden ser similars en contrast amb el seu comportament temporal (on és bastant diferent per cada categoria). Els diferents tipus de programa que classificaven eren anuncis, bàsquet, futbol, notícies i el temps.

Recentment, Glasbert ha presentat una nova aproximació sobre la classificació de seqüències de vídeo MPEG-2 com a *dibuixos animats* o *no dibuixos animats* analitzant el color, la textura i el moviment de frames consecutius en temps real (veure [9] ). En aquest mètode, les característiques extretes dels descriptors visuals s'han ponderat de manera no lineal amb la funció *sigmoid* i després s'ha utilitzat un perceptró multicapa per obtenir un bon reconeixement. Els resultats demostren que l'alta taxa de reconeixement es basa en la llarga base de dades

de 200 seqüències de vídeo (40 de dibuixos animats i la resta de no dibuixos animats) obtingudes gratuïtament de la televisió digital alemanya. Els autors han aconseguit un reconeixement mig del 80% per vídeos de dibuixos animats.

Un altre aspecte a comentar de tots els treballs multimodals fets fins ara és que per dur a terme la classificació s'han seleccionat totes les característiques de manera empírica. Degut al gran nombre de característiques disponibles, un dels aspectes més importants dels classificadors multimodals és que hi ha la necessitat de mesurar la rellevància de les característiques en el procés de classificació, i per tant triar de totes les característiques d'àudio i vídeo les més significatives pels seus propòsits. Això és important per dos motius: el primer, per reduir el cost computacional dels classificadors ja que s'optimitzarien el subconjunt de característiques; i segon, perquè es podria seleccionar de manera eficaç el subconjunt de mesures òptimes depenent del problema que es vol solucionar. És a dir, si es vol detectar dibuixos animats es triarà un conjunt X, si es vol detectar esports, es triarà un altre subconjunt Y. Per aquest motiu un dels classificadors implementats està basat en la Fuzzy Integral ja que segons la teoria (veure 5.4) aquest operador no lineal té la possibilitat de mesurar la rellevància de les característiques, però això s'ha deixat com a línia futura.

Després d'aquest resum sobre alguns dels treballs més representatius sobre la classificació multimodal de gèneres, i més concretament la detecció de dibuixos animats, es descriuen les característiques més importants de l'estàndard MPEG-7 i característiques d'àudio i vídeo que es podrien utilitzar dintre del framework presentat a la Fig. 2.1.

## 2.1. Estàndard MPEG-7

L'objectiu principal de l'estàndard MPEG-7 és descriure les característiques de continguts multimèdia. Proporciona un conjunt de descriptors basats en catàleg (títol, creador, drets), semàntica (informació sobre objectes i events que apareixen en el document) i estructural (com histograma de color) que estandarditza la forma de descriure el contingut audiovisual.

És una tecnologia desenvolupada pel grup MPEG i aprovada per l'Organització Internacional d'Estandardització (ISO) que proporciona una descripció de continguts per paraules clau, significat semàntic (qui, què, quan, on) i estructural (forma, colors, textura i moviments).

El format MPEG-7 s'ha dissenyat perquè sigui independent del format del contingut.

Aquest estàndard s'utilitza per desenvolupar sistemes capaços d'indexar grans bases de dades de material audiovisual i buscar en aquestes bases de dades de manera automàtica. L'estàndard MPEG-7 també inclou informació sobre el tipus de compressió utilitzada (JPEG, MPEG-2 etc.), les condicions per accedir, la classificació, els enllaços a altres materials rellevants i el seu context.

L'estàndard defineix una llibreria multimèdia d'eines i mètodes resumits en la Fig. 2.1.1, i que es detallen a continuació.

*Conjunt de descriptors*

Un descriptor (D) és una representació d'una característica definida de manera sintàctica i semàntica.

*Conjunt d'esquemes de descripció*

Un esquema de descripció (DS) especifica l'estructura i semàntica de les relacions entre els seus components, que poden ser descriptors (D) o esquemes de descripció (DS).

*Llenguatge de descripció definit*

Un llenguatge de descripció (DDL) és un llenguatge que especifica esquemes de descripció permetent l'extensió i modificació dels esquemes de descripcions existents. Es tracta d'un llenguatge basat en XML.

La seqüència desitjada per l'estàndard MPEG-7 és inicialment una anàlisi del document per obtenir les seves característiques i les relacions entre els elements. En aquest cas s'utilitzen els descriptors estàndards que poden ser ampliats.

També es poden utilitzar eines de anotació per descriure la semàntica del document.

D'aquesta manera, utilitzant l'estàndard MPEG-7 qualsevol aplicació podrà entendre i utilitzar la informació obtinguda i serem capaços de desenvolupar potents buscadors o classificadors de documents multimèdia.

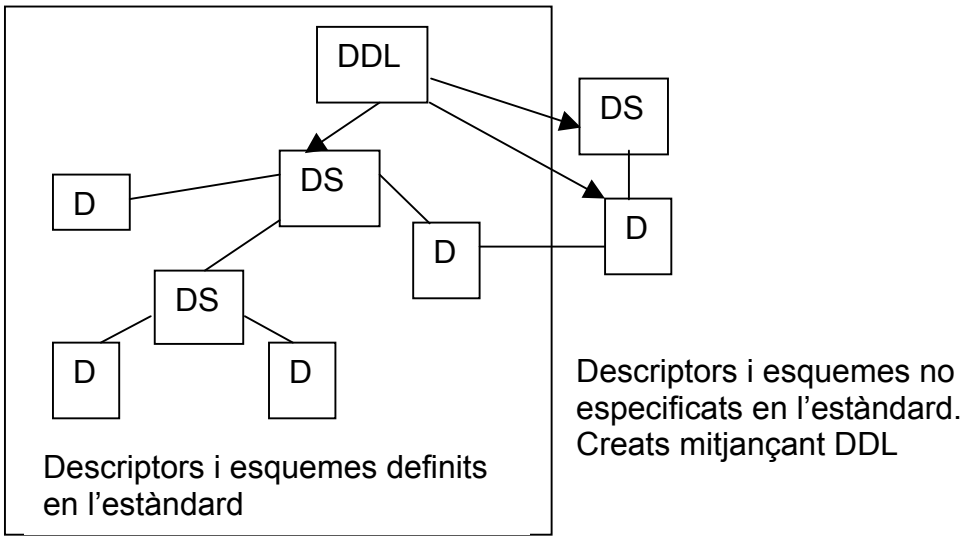



**Fig. 2.1.1** Relació entre elements de l'estàndard MPEG-7

Existeixen moltes aplicacions i àrees on es pot aplicar l'estàndard MPEG-7, ja sigui en biblioteques digitals (catàlegs d'imatges, vídeos etc.), serveis de directoris multimèdia (pàgines grogues, etc.), servei de notícies electròniques, autorització de continguts multimèdia etc.

## 2.2. Àudio

En aquest apartat s'explicaran les característiques per classificar fitxers d'àudio. Per dur a terme una classificació d'àudio hi ha moltes característiques que poden utilitzar-se. Segons l'article "Multimedia Content Analysis Using both Audio and Video Clues" s'extreuen dos nivells de característiques d'àudio: característiques a nivell de trama i característiques a nivell de clip.

Una trama es defineix com un grup de mostres consecutives de 10 a 40 ms de durada (veure [2] ).

Per fer una classificació d' àudio necessitem una anàlisi de duració major a una trama, si s'aprofiten les característiques a nivell de trama i s'apliquen durant un període de temps (1 o 2 segons) es pot fer una anàlisi a nivell de clip.

### 2.2.1. Característiques a nivell de trama

Les característiques a nivell de trama són en un espai curt de temps, d'entre 10 ms i 40 ms. A continuació es defineixen breument cadascuna d'elles.

*Volum*

És la característica més utilitzada i fàcil de calcular. És un indicador del silenci S'aproxima per la rms del senyal en una trama

$$V(n) = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{N-1} S_n^2(i)} \quad \textbf{(2. 1)}$$

on $S_n(i)$ és el valor de la mostra *i* de la trama *n* i N la longitud de la trama.

*Creuaments per zero (ZCR)*

Consisteix en contar les vegades que el senyal d'àudio creua el valor zero en l'eix d'ordenades.

$$ZCR(n) = \frac{1}{2}\left(\sum_{N-1}^{i=1} \left|sign\left(Sn(i)\right) - sign\left(Sn(i-1)\right)\right|\right)\frac{fs}{N} \quad \textbf{(2. 2)}$$

on fs és la freqüència de mostreig i N la longitud de la trama.

*Freqüència fonamental (Pitch)*

És un indicador important en la veu i la música ja que tenen una freqüència fonamental bastant definida. La parla té una freqüència està entre 40 i 450 Hz de freqüència fonamental i la música més alta.

Un mètode per trobar la freqüència fonamental és fer l'autocorrelació i buscar la distància entre pics consecutius.

$$R_n(l) = \sum_{N-l-1}^{i=0} S_n(i) \cdot S_n(i+l) \quad \textbf{(2. 3)}$$

*Freqüència central (FC)*

És una mesura del punt d'equilibri de l'espectre i té una relació directa amb la brillantor del so.

$$FC(n) = \frac{\int_0^\infty w \cdot S_n(w)\, dw}{\int_0^\infty S_n(w)\, dw} \quad \textbf{(2. 4)}$$

on Sn (w) és l'espectre de potència del senyal i w la freqüència angular.

*Ampla de banda (BW)*

És pot calcular a partir del paràmetre anterior FC i la desviació, obtenint una mesura de l'ample de banda del senyal d'àudio.

*Energia en diferents bandes (ESBR)*

Consisteix en la mesura d'energia en diferents subbandes on es divideix l'espectre en quatre bandes amb intervals [0, w/8], [0, w/4], [0, w/2] i [0, w]. La variable *w* representa la meitat de la freqüència de mostreig.

*Spectral Rolloff Point*

Es defineix com el punt de l'espectre on s'assoleix el 95 % de la potència espectral.

*Mel Frequency Ceptrum Coefficients (MFCC)*

S'utilitzen filtres triangulars pas banda espaiats linealment a baixa freqüència i logarítmicament a alta freqüència per extreure les característiques fonètiques importants.

### 2.2.2. Característiques a nivell de clip

Les característiques a nivell de clip són bàsicament mesures estadístiques (mitja o variança) de les anteriors durant una finestra temporal aproximadament d'un o dos segons (veure [2] i [11] ).

## 2.3. Vídeo

En aquest apartat s'expliquen els descriptors de vídeo definits en l'estàndard MPEG-7. Els descriptors per arribar a incloure's a l'estàndard MPEG-7 han de passar uns processos i una sèrie de proves.

Hi ha multitud de descriptors de vídeo i la seva elecció es fa difícil. L'estàndard MPEG-7 proporciona una metodologia utilitzada en la majoria de sistemes de classificació, i per tant la utilització dels seus descriptors permet la interoperatibilitat amb qualsevol altre sistema.

Els descriptors estructurals de MPEG-7 es divideixen en el color, la textura, la forma i el moviment.

### 2.3.1. Color

El color és un dels camps més utilitzats en les característiques d'imatge. Les característiques de color són robustes per observar angles, desplaçaments i rotacions de les regions d'interès. El color és un atribut important pel sistema de visió humana. L'estàndard divideix els descriptors de color en sis descriptors.

*Color Space Descriptor*

Especifica el model de color a utilitzar en un altre descriptor de color, especialment el descriptor de color dominant. Els models de color especificats en el MPEG-7 són RGB, YCbCr, HSV i HMMD.

El model RGB és el més popular. Aquest model es basa en els tres colors primaris (vermell, verd i blau) per representar els altres colors.

El model YcbCr és una transformació lineal del RGB i s'expressa com:

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$
$$Cb = -0.169 \cdot R - 0.331\, G + 0.500 \cdot B$$
$$Cr = 0.500 \cdot R - 0.419 \cdot G - 0.0081 \cdot B \quad \textbf{(2. 5)}$$

El model HSV es defineix com un cilindre amb les variables tonalitat (H), la intensitat (V) i la saturació (S). La tonalitat es representa per un angle que va de 0 a 360º i especifica la família del color. La intensitat especifica la brillantor o foscor del color i la saturació la puresa del color.

El HMMD (Hue-Max-Min-Diff) és un model de forma cònica on els noms Hue, Max, Min i Diff segueixen les equacions següents:

$$Max = \max\left(R, G, B\right)$$
$$Min = \min\left(R, G, B\right)$$
$$Diff = Max - Min$$
$$Sum = \frac{Max + Min}{2} \quad \textbf{(2. 6)}$$

*Dominant Color Descriptor (DCD)*

És un descriptor que proporciona una informació compacta dels colors representatius d'una imatge o regió de la imatge.

*Scalable Color Descriptor*

És un descriptor que es pot interpretar com la transformada Haar utilitzant el mapa de colors HSV .

*Group of Frames Descriptor (GoF)*

És un descriptor utilitzat per representar característiques de color de múltiples imatges o múltiples frames.

*Color Structure Descriptor*

Representa una imatge amb la seva distribució de color (similar al histograma de color) i l'estructura espaial de color utilitzant una estructura de finestres.

*Color Layout Descriptor*

Captura els colors representatius en una quadrícula. La representació es basa en els coeficients de la transformada cosinus discreta. S'utilitza tant en imatges com en segments de vídeo.

### 2.3.2. Textura

La textura es refereix a patrons visuals que tenen propietats d'homogeneïtat o no, i la presència de múltiples colors o intensitats en la imatge. L'estudi de la textura d'una imatge ha sorgit com a conseqüència de buscar a través de llargues col·leccions d'imatges de textura similar.

Molts objectes naturals i artificials es poden distingir per la seva textura. Una imatge es pot considerar com un mosaic de textures homogènies.

Actualment hi ha tres descriptors de textura en l'estàndard MPEG-7: el descriptor homogeni de textura (HTD), el descriptor buscador de textura (TBD) i l' histograma de contorn (EHD).

*Texture Browsing Descriptor*

El TBD especifica la caracterització perceptual de la textura, en termes de regularitat, direcció i intensitat.

*Edge Histogram Descriptor*

El EHD consisteix en dividir la imatge en 4x4 imatges i generar un histograma de cada una de les imatges de cinc maneres diferents: vertical, horitzontal, 45º diagonal, 135º diagonal i sense direcció.

*Homogeneous Texture Descriptor*

LA HTD és el descriptor implementat en aquest projecte i s'explica en més detall en el 4.1.

### 2.3.3. Forma

La forma és una característica important de baix nivell. Els descriptors de forma es poden dividir en descriptors basats en el contorn i descriptors basats en la regió.

Els mètodes basats en el contorn exploten la informació frontera que és crucial en la percepció humana per la detecció de formes similars.

Els mètodes basats en la regió exploten només informació interior, que pot ser aplicada a formes més generals.

### 2.3.4. Moviment

Els descriptors MPEG-7 de moviment són clau en la indexació de vídeo, són descriptors compactes i representatius.

L'estàndard consta de quatre descriptors: l'activitat de moviment, el moviment de la càmera, la trajectòria de la càmera i el moviment paramètric.

Els moviments de càmera i l'activitat de moviment descriuen característiques de moviment en segments de vídeo, mentre que la trajectòria de moviment descriu el moviment en regions. El moviment paramètric pot ser utilitzat ambdós casos.

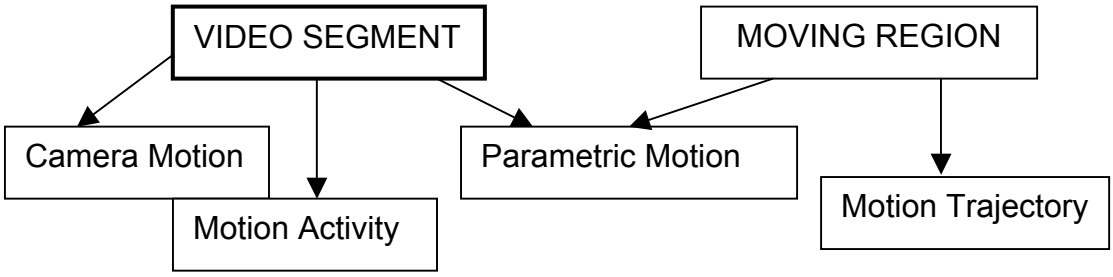



**Fig. 2.3.1** Descriptors de moviment

*Motion Activity*

Aquest descriptor consta de quatre atributs: la intensitat de l'activitat, la direcció de l'activitat, la distribució espacial i la distribució temporal.

*Camera Motion*

Camera Motion suporta totes les operacions de la càmera com les translacions, rotacions, canvi de focalització i combinacions d'aquestes.

Dos segments de vídeo de la mateixa escena amb diferents moviments de càmera porten a diferents representacions de la realitat.

*Motion Trajectory*

Descriu la trajectòria dels objectes en temps, definint objecte com una regió espai-temps.

*Parametric Motion*

Representa el moviment i deformació d'una regió o imatge a través de models paramètrics (translació, rotació, perspectiva, quadràtic).

# CAPÍTOL 3. FITXERS DE CARACTERÍSTIQUES

En aquest capítol s'explicarà la metodologia seguida en aquest projecte des de tenir un fitxer de vídeo fins a l'obtenció d'un fitxer de característiques del vídeo analitzat. Tal com es mostra en la Fig.0.2 , comentarem les cinc primeres etapes: gravació audiovisual, segmentació manual d'escenes, etiquetatge manual, extracció de característiques i obtenció dels macrofitxers .

La utilitat d'obtenir un fitxer de característiques és poder entrenar un classificador que permeti distingir el gènere de dibuixos animats dels altres gèneres (esports, notícies, documentals etc.).

## 3.1. Segmentació manual

S'ha disposat de dos fitxers de vídeo gravats directament de TV3 en format VOB. Els dos fitxers de vídeo tenen una temàtica de dibuixos animats i d'una duració aproximada de 40 minuts cadascun. S'ha realitzat una conversió de format de VOB a MPEG-2 amb les següents característiques:

| | |
|---|---|
| Resolution: | 720x576 |
| Aspect Ratio: | 4:3 |
| Frame Rate: | 25fps – PAL (625/50) |
| Video Bit Rate: | 6Mbps (constant) |
| Audio Bit Rate: | 384Kbps |

Per tal de facilitar l'extracció de les característiques del vídeo és necessària una segmentació prèvia del vídeo en shots (escenes). El programa utilitzat per la segmentació manual ha estat l'Adobe Premiere 6.5 amb el LSX-MPEG Suite 2.0 (Premiere Plug in).

**Fig. 3.1.1** Adobe Premiere 6.5

S'ha intentat que la fragmentació fos coherent (una mateixa escena, una conversa, ...) pensant sempre en el posterior anàlisi.

Finalment, s'han obtingut 80 fragments de dibuixos animats de poca duració com a resultat de la segmentació dels dos vídeos inicials.

## 3.2. Etiquetatge manual

Per tal de disposar una bona base de dades de fragments de vídeo per a una posterior utilització és necessari l'etiquetatge de cada fragment. S'han seguit unes premisses per fer l'etiquetatge amb un contingut com el de la Fig. 3.2.1.

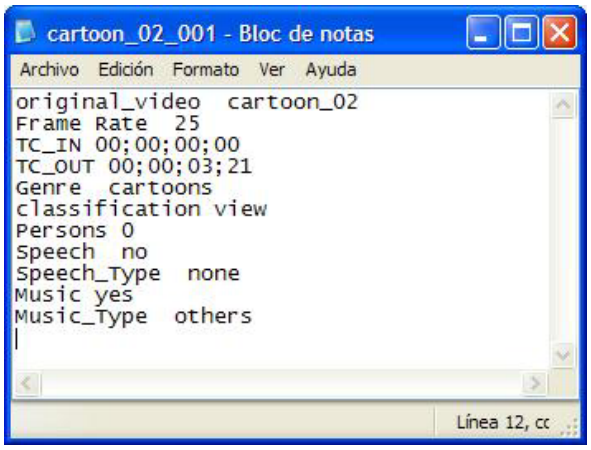



**Fig. 3.2.1** Etiquetatge dels fragments de vídeo

## 3.3. Exportació dels fitxers

Seguint l'esquema de la Fig. 2.1 s'ha exportat de cada fragment un fitxer d'àudio (.wav) i un de vídeo (.avi) utilitzant també el Adobe Premiere 6.5.

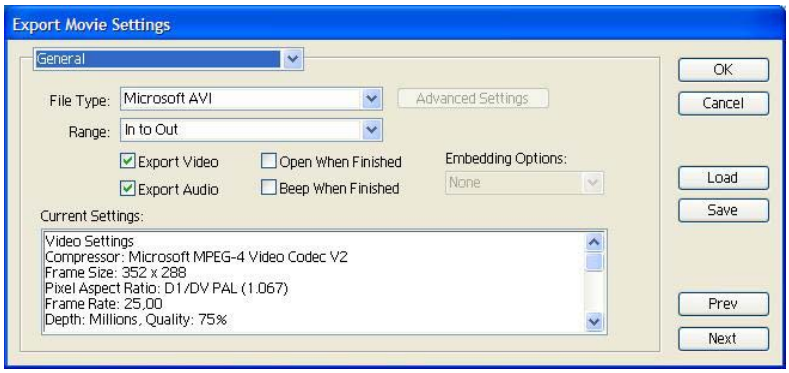



**Fig. 3.3.1** Exportació del fitxer de vídeo

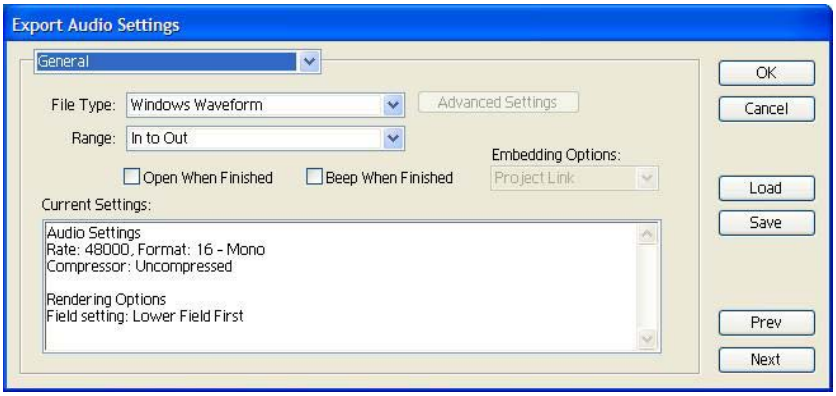



**Fig. 3.3.2** Exportació del fitxer d'àudio

Una característica important sobre els fitxers de vídeo és que les dimensions s'han reduït a 352 x 288 píxels. Aquestes dimensions seràn les utilitzades durant tot el projecte.

## 3.4. Extracció de característiques

Els fitxers exportats (.avi i .wav) s'introdueixen a un programa que extreu les característiques d'àudio i vídeo. Aquestes característiques s'emmagatzemen en un fitxer de text. Cada fragment genera un fitxer de text amb valors numèrics per cada frame, és a dir, gran contingut d'informació.

## 3.5. Generació dels fitxers d'entrenament

Per facilitar l'entrenament del classificador és interessant que no s'hagin de manipular tots els fitxers de text generats per l'aplicació de l'apartat anterior d'un manera individual. Per tant, es crea un únic fitxer que engloba tots els altres i fa més àgil l'entrenament del classificador.

# CAPÍTOL 4. CARACTERÍSTICA DE TEXTURA

El treball d'estandardització MPEG-7 va començar amb l'objectiu de proveir eines fonamentals per descriure el contingut multimèdia. En aquest capítol parlarem d'un descriptor de textura anomenat HTD (homogeneous texture descriptor). El descriptor HTD actual consisteix en mitja, desviació estàndard de la imatge i energia, desviació energètica de la transformada de Fourier de la imatge.

Aquests són extrets a partir de dividir canals de diferents freqüències basats en el sistema visual humà (HVS). Per assegurar l'extracció del descriptor de textura s'utilitza la transformació Radon. És aconsellable pel comportament del sistema visual humà.

La informació de textura d'una imatge és una característica fonamental, en el qual s'ha estudiat durant dècades per analitzar imatges en el món de la medicina, imatges per satèl·lit...

Treballs anteriors com la distribució probabilística de píxels, filtres direccionals i camps aleatoris de Markov han estat estudiats (veure [5] ). Més recentment, s'han estudiat filtres Gabor i transformades wavelet per extreure informació de textura.

El MPEG-7 HTD és eficient no només per processar informació de textura, sinó per representar la informació. Consisteix en la mitja i desviació estàndard de la imatge. També inclou l'energia i la desviació energètica de la transformada de Fourier de la imatge.

## 4.1. Homogeneous Texture Descriptor

### 4.1.1. Sistema de visió humana (HVS)

Últimament, tècniques de caracterització de textura basades en HVS han estat proposades. En aquests experiments, s'observa que la resposta de la cortesa visual està en una banda limitada en el domini freqüencial. El cervell humà divideix l'espectre en canals que són bandes en l'espai freqüencial. Per caracteritzar la textura, la millor representació de sub-bandes de l'HVS és la divisió de l'espai freqüencial en octaves (4-5 divisions) a través de la direcció radial i amb igual amplada en direcció angular.

Aquestes sub-bandes són simètriques respecte l'origen de coordenades polars. Aquesta estructura freqüencial permet extreure informació de textura d'una manera similar al sistema visual humà. En aquestes sub-bandes s'extreuen les característiques de textura.

L'espai freqüencial d'on s'extreuen les característiques es divideix en la direcció angular en angles de 30º i en la direcció radial en octaves. Les sub-bandes s'anomenen canals, per tant l'espai freqüencial està dividit en 30 canals com es veu en la Fig. 4.1.1.

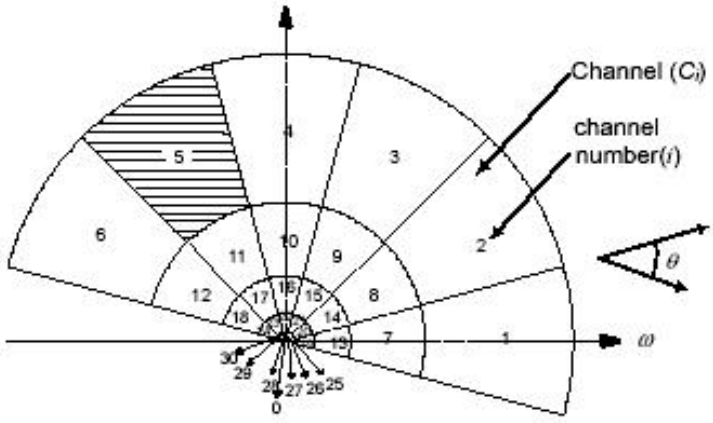



**Fig. 4.1.1** Canals utilitzats en la HTD

L'espai freqüencial està normalitzat $(0 \le w \le 1)$ , on w és $w = \Omega / \Omega_{max}$ . Les freqüències centrals dels canals estan espaiades 30º. $\theta r$ segueix la formula següent $\theta r = 30^o \cdot r$ , on r és l'índex angular amb $r \ \varepsilon \ \{0,1,2,3,4,5\}$ . En la direcció radial, les freqüències centrals de cada canal estan espaiades en octaves seguint aquesta fórmula $Ws = Wo \cdot 2^{-S}$, on s és l'index radial $s \ \varepsilon \ \{0,1,2,3,4\}$ i $Wo$ té un pes de 3/4. L'ample de banda dels canals en la direcció radial s'escriu com $Bs = Bo \cdot 2^{-S}$ , on Bo és l'ample de banda més alt amb valor 1/2.

Com es veu en la Fig. 4.1.1 cada regió correspon a una porció de l'ample de banda que és la resposta de la cortesa visual humana. Els canals situats a baixes freqüències tenen una mida més petita, mentre que a altes freqüències l'àrea és més gran. Això correspon a la visió humana que és més sensible a canvis a baixa freqüència. S'assumeix d'estar sempre en valors reals, per això estem en la meitat del pla freqüencial.

## 4.1.2. Descripció del descriptor

Amb les divisions de la Fig. 4.1.1, les regions freqüencials són més petites comparades amb les d'alta freqüència en el sistema de coordenades cartesianes, per tant les mostres a baixa freqüència són escasses on la informació de textura és insuficient.

Per solucionar això s'utilitza la transformada Radon a la imatge, en el qual permet a la transformada de Fourier d'una imatge en coordenades cartesianes

ser representada en coordenades polars. Utilitzant la transformada Radon, una imatge 2D pot transformar-se en una projecció d'una dimensió, l'espai cartesià (x ,y) és pot representar en un espai Radon $(R,\theta)$ segons la Fig. 4.1.2.

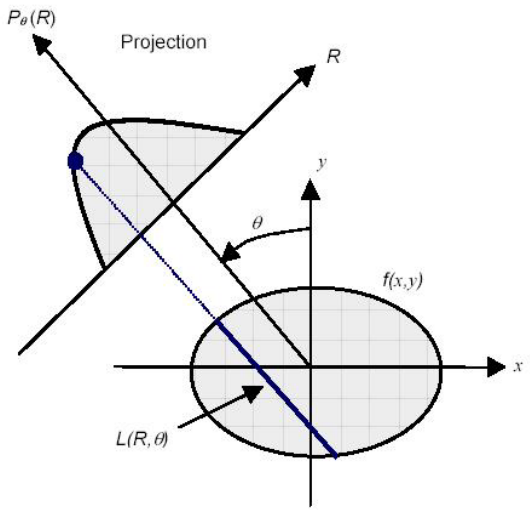



**Fig. 4.1.2** Transformada Radon. La imatge (x,y) es transforma en P(R) en l'espai Radon $(R,\theta)$

L'integral de línia a través de la línia $L(R,\theta)$ en un angle $\theta$ en el sentit de les agulles del rellotge des de l'eix y i a una distància R de l'origen es pot escriure com

$$\mathrm{P}\theta\,(R) = \int_{L(R,\,\theta)} f(x\,,y)\;dl = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x\,,y)\,\delta\,\left(x\cos\,\theta + y\sin\theta - R\right)dxdy \quad \textbf{(4. 1)}$$

on f ( x, y) és la funció de la imatge, R és l'eix de projecció i $\delta$ (.) és la funció delta. La funció P$\theta$ ( R ) és una projecció. El recull complet d'integrals de línia s'anomena transformada Radon de f ( x , y ) i també s'anomena Sinograma. Les propietats freqüencials en la transformada Radon es poden explicar a través de "central slice theorem" en el qual la projecció d'una transformada de Fourier d'una dimensió amb angle $\theta$ és igual a un tall d'angle $\theta$ a través de la transformada de Fourier 2D d'una imatge (veure Fig. 4.1.3).

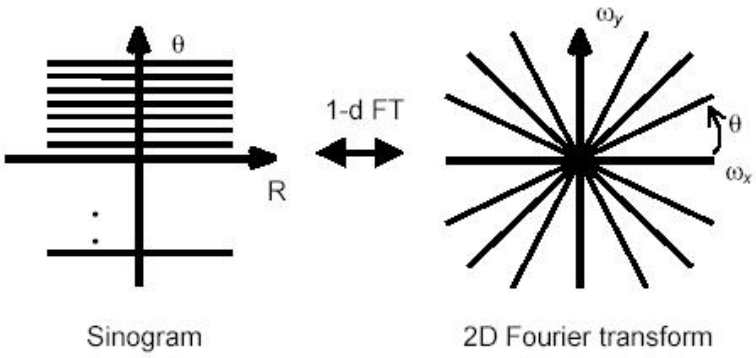



**Fig. 4.1.3 Relació entre un sinograma i la transformada de Fourier 2D**

La projecció de la transformada de Fourier d'una dimensió es pot escriure com

$$\int P\theta\,(R)\, e^{-j\,2\Pi R\,w}\; dR = \int\int f(x\,,y)\, e^{-j\,2\Pi\,w\,(\,x\,\cos\theta\,+\,y\,\sin\theta\,)}\; dxdy \qquad \text{on}$$

$$\sqrt{Wx^2 + W\,y^2} \quad \text{i} \quad \theta = \tan^{-}1\left(\frac{Wy}{Wx}\right) \qquad \textbf{(4. 2)}$$

La Fig. 4.1.4 mostra l'estructura en coordenades polars després de la transformada Radon i de la seva transformada de Fourier. La densitat freqüencial és més densa a baixa freqüència. Aquesta propietat correspon a la visió humana.

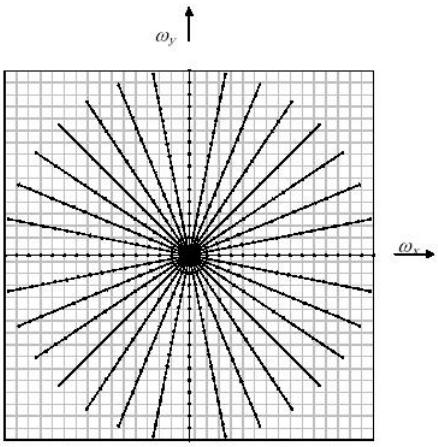



**Fig. 4.1.4** Domini freqüencial després de la trasformada Radon i la seva transformada de Fourier

La Fig. 4.1.1 s'implementa a través de filtres ideals que agrupen les característiques de cada canal. Els filtres Gabor es defineixen com

$$Gp\,s,r\left(w,\theta\right) = e^{-\frac{(w-ws)^2}{2\sigma ws^2}}\; e^{-\frac{(\theta-\theta r)^2}{2\sigma\theta r^2}} \qquad \textbf{(4. 3)}$$

on Gp s,r és la funció en s (índex radial) i r (índex angular). La Fig. 4.1.5 mostra els filtres Gabor, on hi han 6 divisions en la direcció angular i 5 divisions en la direcció radial.

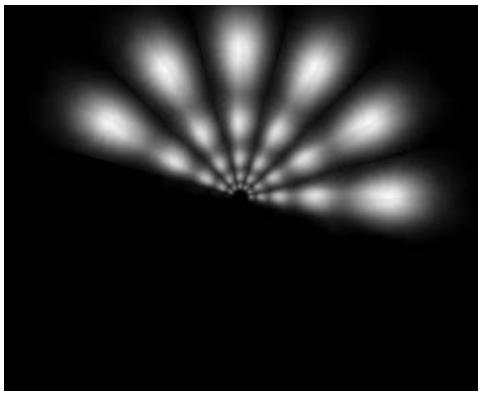


**Fig. 4.1.5** Filtre Gabor 6x5

La $\sigma_{\theta r}$ és una constant de valor $\frac{15^o}{\sqrt{2\ln 2}}$ en la direcció angular. En la direcció radial $\sigma_{ws}$ depèn de l'ample de banda i s'escriu

$$\sigma_{Ws} = \frac{Bx}{2\sqrt{2\ln 2}} \qquad \textbf{(4. 4)}$$

Les següents taules mostren les característiques dels canals de la funció Gabor.

**Taula. 4.1.1** Paràmetres en la direcció radial

| Index radial (s) | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Freqüència central (Ws) | 3 / 4 | 3 / 8 | 3 / 16 | 3 / 32 | 3 / 64 |
| Ample de banda (Bx) | 1 / 2 | 1 / 4 | 1 / 8 | 1 / 16 | 1 / 32 |
| $\sigma_{ws}$ | $\frac{1}{4\sqrt{2\ \ln 2}}$ | $\frac{1}{8\sqrt{2\ \ln 2}}$ | $\frac{1}{16\sqrt{2\ \ln 2}}$ | $\frac{1}{32\sqrt{2\ \ln 2}}$ | $\frac{1}{64\sqrt{2\ \ln 2}}$ |

**Taula. 4.1.2** Paràmetres en la direcció angular

| Index angular (r) | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Freqüència central ($\theta$r) | 0º | 30º | 60º | 90º | 120º | 150º |
| Ample de banda | 30º | 30º | 30º | 30º | 30º | 30º |
| $\sigma_{\theta r}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ | $\frac{30^o}{2\sqrt{2\ \ln 2}}$ |

Per extreure les característiques, agafem la transformada Radon de la imatge i fem la transformada de Fourier d'una dimensió, Després, es pot obtenir un tall central de F(w, $\theta$) en domini freqüencial 2D. El descriptor de textura consisteix en les característiques de cada canal. Els components són el primer i segon mo ment d'energia en els canals (desviacions energètiques). Les energies i desviacions energètiques s'escriuen com $[e_1, e_2, e_3, \ldots, e_{30}]$ i $[d_1, d_2, d_3, \ldots, d_{30}]$ respectivament, Els índexs indiquen els canals. L'energia *ei* d'un canal *i* es defineix com

$$e_i = \log\left(1 + p_i\right) \quad \textbf{(4. 5)}$$

$$pi = \sum_{w=0}^{1} \sum_{\theta=0}^{360} \left[ G_{p\,s,r}\left(w,\theta\right) \cdot |w| \cdot F\left(w,\theta\right) \right]^2 \quad \textbf{(4. 6)}$$

$$d_i = \log\left(1 + q_i\right) \quad \textbf{(4. 7)}$$

$$q_i = \sqrt{\sum_{w=0}^{1} \sum_{\theta=0}^{360^o} \left\{ \left[ G_{P_{S,R}}\left(w,\theta\right) \cdot |w| \cdot F\left(w,\theta\right) \right]^2 - p_i \right\}^2} \quad \textbf{(4. 8)}$$

on |w| és el terme jacobià entre la freqüència polar i la cartesiana i es pot escriure com $|w| = \left| \sqrt{w_x^2 + w_y^2} \right|$ . F (w, $\theta$ ) és la transformada de Fourier de la imatge f(x,y).

Finalment, el descriptor bàsic consisteix en un vector de 32 valors on els dos primers són la mitjana i la desviació típica, i la resta les 30 energies dels canals.

$$TD = \left[ f_{DC}, f_{SD}, e_1, e_2, e_3, \ldots, e_{30} \right] \quad \textbf{(4. 9)}$$

També existeix el descriptor avançat on s'inclouen les desviacions energètiques.

$$TD = \left[ f_{DC}, f_{SD}, e_1, e_2, e_3, \ldots, e_{30}, d_1, d_2, d_3, \ldots, d_{30} \right] \quad \textbf{(4. 10)}$$

# CAPÍTOL 5. CLASSIFICADORS

En aquest apartat explicarem breument els classificadors utilitzats per la classificació de gènere de dibuixos.

Inicialment els classificadors necessitaran una fase d'entrenament on s'ajustaran els seus pesos, i posteriorment es calcularà la probabilitat que els vídeos pertanyin al gènere de dibuixos animats.

S'ha implementat un total de 4 classificadors: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Mean-Nearest Neighbour (M-NN) i un classificador no lineal basat en la Fuzzy Integral.

És necessari aclarir que tant la PCA com la LDA no són estrictament uns classificadors. Aquests dos mètodes calculen un espai transformat a partir d'un conjunt de mostres d'entrenament (fase d'entrenament) segons un criteri de maximització de l'energia. És a dir, s''obté una codificació més compacta dels objectes que classificarem (dibuixos i no-dibuixos) facilitant i millorant els resultats.

A continuació s'explicarà amb més detall cadascuna d'aquests classificadors, particularitzant pel reconeixement de dos classes: dibuixos i no dibuixos.

## 5.1. PCA

Aquest mètode consisteix en reduir la dimensió de l'espai de representació dels objectes a partir de projeccions lineals.

La manera d'operar és la següent, sigui N objectes representats per vectors de característiques $\{X_1, X_2, \ldots, X_N\}$ tals que $x_i \varepsilon R^n$ , considerem una transformació lineal que transforma l'espai orignial n-dimensional en un espai m-dimensional , sent m<n. Les noves representacions $y_i$ es calculen

$$y_i = W^T \times x_i \quad i = 1,2, \ldots, N \quad \textbf{(5. 1)}$$

sent les *m* les columnes de $W \varepsilon R^{n \times m}$ ortonormals. Sigui $S_c$ la matriu de covariància dels vectors de característiques originals definida com

$$S_c = \left(\frac{1}{N}\right) \times \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \quad \textbf{(5. 2)}$$

on $\mu \varepsilon R^n$ representa la mitja dels vectors de característiques originals. $S_c$ té dimensions n x n, és simètrica i té n vectors propis $\{\Phi_1, \Phi_2, \ldots, \Phi_n\}$ i n valors propis $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ que compleixen la relació

$$S_c \times \Phi = \Phi \times \Lambda \quad \textbf{(5. 3)}$$

on,

$\Phi$ és la matriu de vectors propis $\{\Phi_1, \Phi_2, \ldots, \Phi_n\}$

$\Lambda$ és la matriu de valors propis $\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$

els $\Phi_i$ són ortogonals entre si, per tant la matriu $\Phi$ es defineix una nova base, sense deformar l'espai original

$$y_i = \Phi^T \times x_i \quad \textbf{(5. 4)}$$

després d'aquesta transformació els vectors $y_i$ tenen una matriu de covariàncies

$$\Phi^T \times S_c \times \Phi \quad \textbf{(5. 5)}$$

ja que $\Phi$ és ortogonal $\left(\Phi^T \times \Phi = I\right)$ i utilitzant la (5.3), la nova matriu de covariància és $\Lambda$ . Amb aquesta transformació s'ha aconseguit decorrelar els valors de les noves dimensions.

Un exemple d'aquesta transformació pot veure's en la Fig. 5.1.1.

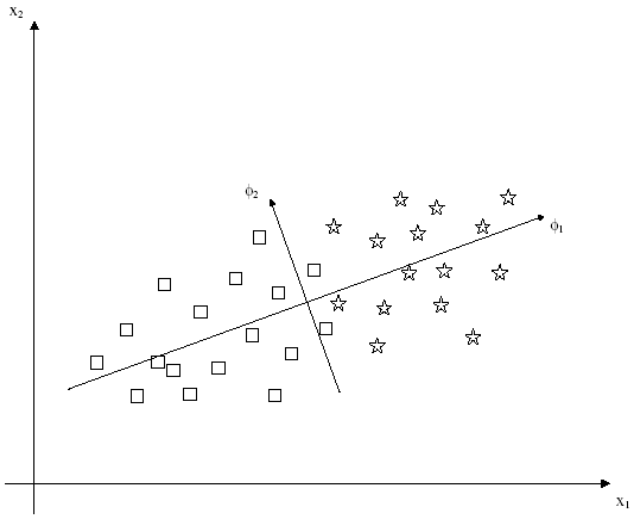



**Fig. 5.1.1** Aplicació PCA

Particularitzant en el classificador utilitzat en el projecte, s'escull la mitja del vector de cada classe més representativa i s'obté un vector de característiques per cada classe. Aquest vector es compara amb el vector de test ($\bar{f}_{TEST}$) a través de la distància mètrica. La distància més pròxima és la classe reconeguda (dibuixos o no dibuixos).

La Fig. 5.1.2 mostra l'estructura d'entrenament i reconeixement del classificador PCA utilitzat en el projecte.

FASE ENTRENAMENT FASE RECONEIXEMENT

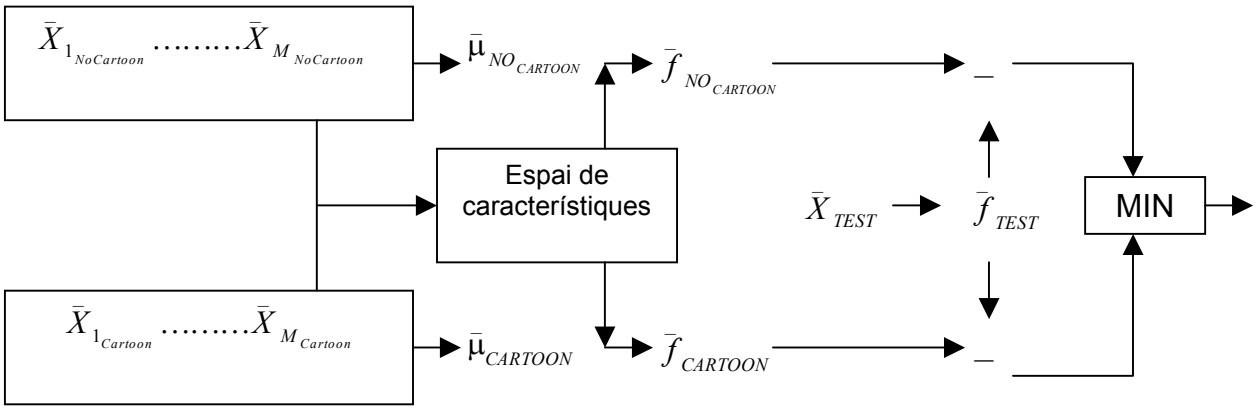



**Fig. 5.1.2** Estructura del PDA

## 5.2. LDA

En aquest classificador la diferència entre classes (dibuixos o no dibuixos) es maximitza, mentre la dispersió de les mostres de cada objecte es minimitza. És a dir, consisteix en trobar la direcció en la qual les mostres del conjunt d'entrenament tenen la mínima dispersió dins d'una classe i la màxima dispersió entre classes i projectar la mostra a classificar sobre aquesta direcció.

En l'anàlisi discriminant lineal també es busca una matriu W que defineixi una transformació lineal. El LDA utilitza la informació de classe.

Per fer ús de la informació de classe es defineixen la matriu de covariància $\left(S_b\right)$ i la matriu de covariància $\left(S_w\right)$. La matriu de covariàncies es defineixen com

$$S_b = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right) \times \left(M_0 - M_i\right) \times \left(M_0 - M_i\right)^T \quad \textbf{(5. 6)}$$

on *L* és el nombre de classes, *N* el nombre total de vectors, *Ni* el nombre total de vectors de la classe *i*, *Mo* és el vector mitja dels vectors i *Mi* és el vector mitja dels vectors de la classe *i*.

La matriu de covariància es defineix com

$$S_w = \sum_{i=1}^{L} \left(\frac{N}{N_i}\right) \times S_i \quad \textbf{(5. 7)}$$

on Si és la matriu de covariància de la classe i que es calcula com

$$S_i = \left(\frac{1}{N_i}\right) \times \sum_{x \varepsilon C_i} \left(M_i - x\right) \times \left(M_i - x\right)^T \quad \textbf{(5. 8)}$$

Ci és el conjunt de vectors de la classe i.

Amb aquestes matrius podem definir el criteri de separabilitat $J = tr\left(S_w^{-1} \times S_b\right)$ (criteri de Fisher). Com més gran sigui la distància entre classes i menor la classe, major serà J i més fàcil serà la classificació. Altres criteris com $J_2 = tr\left(S_c^{-1} \times S_b\right)$ o $J_3 = tr\left(S_w^{-1} \times S_c\right)$ també són possibles.

On $\left\{\Phi_i : 1 \ldots L-1\right\}$ són els L-1 vectors propis amb valor propi diferent de zero de la matriu $S_w^{-1} \times S_b$ .

En la Fig. 5.2.1 s'observa un exemple on la PCA ens donaria una transformació que augmentaria el solapament entre classes, mentre que la LDA les separaria correctament.

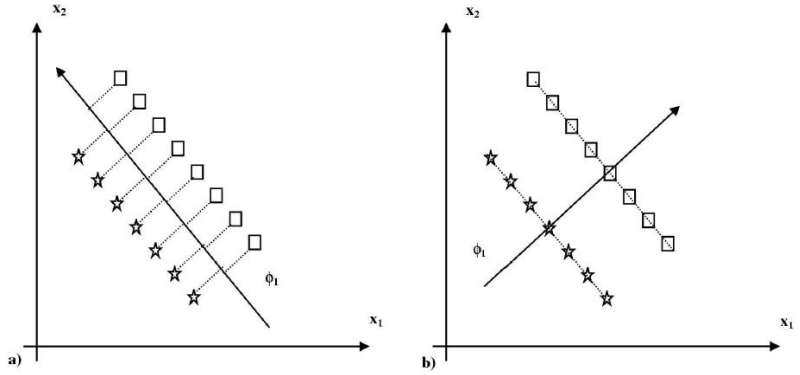



**Fig. 5.2.1** a) Projecció PCA. b) Projecció LDA

## 5.3. M-NN

És un classificador que es basa en la comparació del vector d'entrada respecte el vector de referència o entrenament.

En l'entrenament la mitjana de totes les mostres corresponents a una classe són emmagatzemades i utilitzades com a vector de referència per aquesta classe.

En la fase de reconeixement, es calcula la distància Euclídea entre la mostra a classificar i la mitjana de cadascuna de les classes a reconèixer (dibuixos i no dibuixos en el nostre cas). Es selecciona la classe de la mostra que està a una distància menor. En la Fig. 5.3.1 es mostra l'estructura d'aquest classificador.

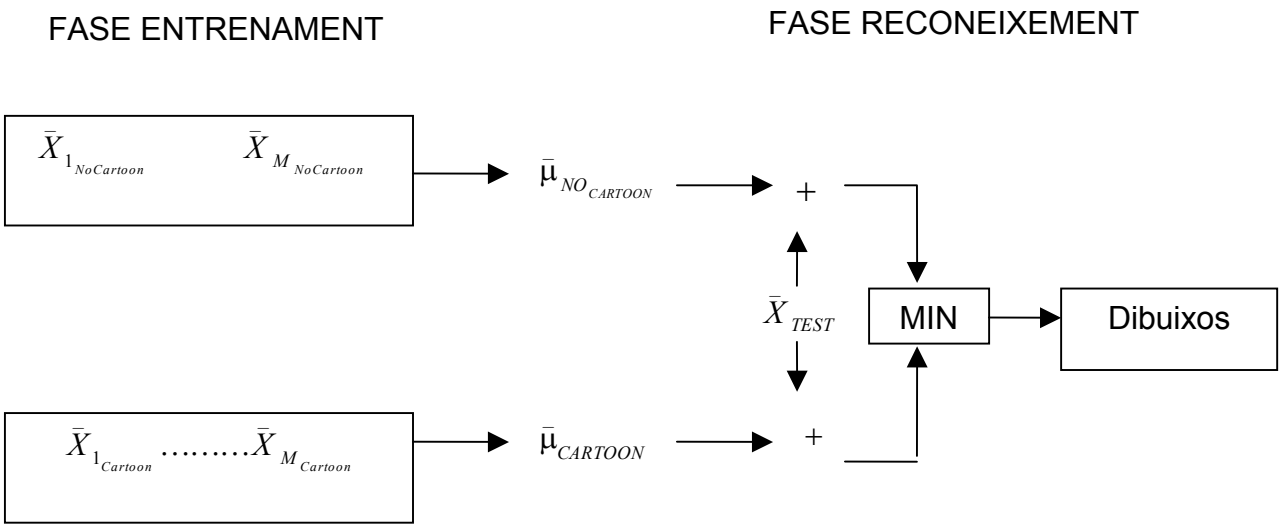



**Fig. 5.3.1 Classificador M-NN**

## 5.4. Fuzzy Integral

La integral difusa (Fuzzy Integral) és un operador no-lineal de fusió introduït per Sugeno a l'any 1974. Es tracta d'un operador utilitzat per resoldre problemes en sistemes automatitzats de presa de decisions multicriteri, avaluació subjectiva multiatribut i fusió d'informació pel reconeixement de patrons (veure [10]).

La idea intuïtiva és la simulació del procés d'integració de diferents fonts d'informació per part de l'ésser humà. Cada persona pren una decisió basant-se en diferents atributs, i després realitza una ponderació segons les seves prioritats.

Aquest procés d'assignació és complex ja que s'han de valorar diferents criteris de forma individual i les combinacions d'aquestes. El procés de síntesi final no es fa a través d'una suma ponderada, sinó que es fa a través de la integral difusa.

Existeixen dos tipus d'integrals difuses, la de Sugeno i la de Choquet. La seva diferència és el tipus d'operadors implicats en la fusió de les dades, i que en la teoria de conjunts difusos reben el nom de normes T i S. En aquest projecte s'utilitzarà la integral difusa de Choquet que segueix la formula

$$\int f \times d\mu = \sum_{i=1}^{n} \left[ f\left(x'_i\right) - f\left(x'_{i-1}\right) \right] \times \mu\left( \left\{ x'_i, x'_{i+1}, K \times x'_n \right\} \right) \qquad \textbf{(5. 9)}$$

on $\mu\left(\left\{x'_i, x'_{i+1}, K \times x'_n\right\}\right)$ són les mesures difuses i $f\left(x'_i\right)$ els vectors dels valors de les característiques.

Aquest classificador també té la capacitat de mesurar la rellevància de les característiques que s'estan avaluant, i per tant determinar quines són les característiques a utilitzar en cada moment per a la classificació.

# CAPÍTOL 6. PROGRAMES

En aquest capítol explicarem les eines utilitzades, l'estructura i el funcionament dels dos programes elaborats.

## 6.1. MacroFitxer

És una eina que facilita la manipulació de les característiques extretes de cada segment de vídeo unificant les dades en un sol fitxer de text. Aporta agilitat i un valor afegit alhora d'entrenar el classificador.

S'ha utilitzat el Visual Studio 2003 com a entorn de programació, i com a llenguatge el C++.

El què fa aquesta classe és a partir d'un directori comprovar tot el seu contingut i buscar els fitxers que acabin amb *_s_VideoFeatures.txt. Aquests són els fitxers de texts corresponents a les característiques de cada segment de vídeo.

Els fitxers es van obrint consecutivament, les dues primeres línies de cada fitxer de text s'eliminen i les altres es van copiant en un nou fitxer.

### 6.1.1. Funcionament

La interfície gràfica és simple, s'ha de seleccionar el fitxer i prémer "Acceptar". En aquest directori hi ha d'haver tots els altres fitxers de característiques que es desitgin adjuntar en el nou fitxer.

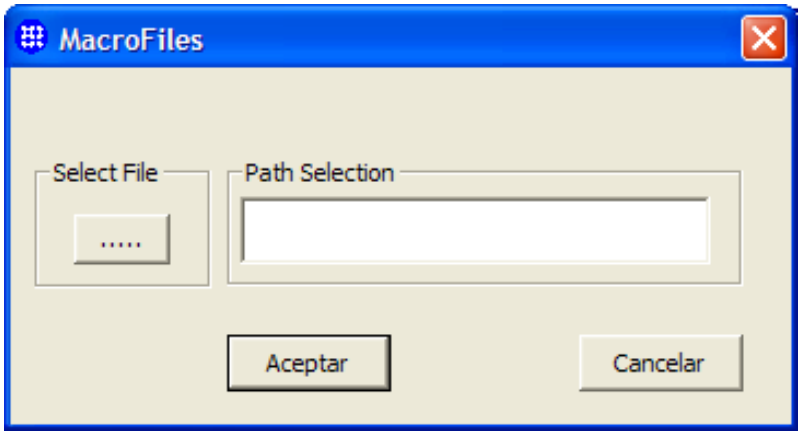



**Fig. 6.1.1** Aplicació per generar un macro fitxer

És un programa fet a mida pels fitxers de característiques de textura, per tant si les característiques dels fitxers a adjuntar són de diferent amplada, el fitxer generat serà erroni.

## 6.2. Descriptor de Textura

### 6.1.1. Eines utilitzades

Per realitzar aquest descriptor de textura l'entorn de programació i el llenguatge utilitzats han estat els mateixos que en l'altre programa (Visual Studio 2003, i C++). També s'han utilitzat simultàniament tres llibreries de processat d'imatge.

*OpenCV*

OpenCV (Open Source Computer Vision Library) és una llibreria oberta desenvolupada per Intel. Aquesta llibreria proporciona un alt nivell de funcions pel processat d'imatge. Aquestes llibreries permeten crear aplicacions poderoses en el domini de la visió digital. OpenCV és opensource per poder funcionar amb moltes plataformes.

Permet fer operacions bàsiques, processat d'imatge, anàlisi estructural, anàlisi de moviment, reconeixement del model, reconstrucció 3D i calibració de la càmera, interfície gràfica i adquisició...

És compatible amb Intel Processing Library (IPL) i utilitza l'estructura IplImage per crear i manipular imatges. Aquesta estructura té molts camps, però en aquest projecte només s' utilitzen alguns d'ells (width, height, depth, nChannels i widthstep ). També disposa d' una interfície gràfica anomenada highGUI.

No obstant, té alguns inconvenients com són la necessitat d'utilitzar una llibreria IPL per tenir funcions de més baix nivell.

Des del seu llançament al 2000, OpenCV ha vist més de 5000 membres registrats al seu grup d'usuaris.

*IPL (Intel Image Procesing Library)*

És una llibreria d'Intel pel tractament d'imatges que consta d'un conjunt de funcions a baix nivell en format DLL. Aquestes funcions estan optimitzades per processadors Intel.

Són funcions optimitzades i per tant es redueix el temps d'execució i permet realitzar experiments més complexes.

És una llibreria de domini públic i està desenvolupada per Intel. Intel garantitza el seu suport i futures actualitzacions.

La llibreria conté funcions aritmètiques, filtres, transformacions geomètriques i lineals, conversions de models de color, histogrames i estadístiques sobre les imatges.

La utilització d'aquesta llibreria és exclusivament per entorn Windows, i s'ha utilitzat per entorns Microsoft Visual C++, Borland C++ i Delphi.

*IPP (Intel Integrated Performance)*

És una llibreria multiplataforma que permet escriure aplicacions per maximitzar el rendiment en processadors Intel.

IPP inclou funcions de processat d'imatge i senyal, xifrat, manipulació de cadenes de text i vectors.

En aquest projecte s'han utilitzat les IPPi que són les llibreries de processat d'imatge.

Ha estat de gran ajuda el fòrum que disposa Intel sobre aquestes llibreries per poder fer consultes en la utilització del seu software.

*Video codec MPEG-4*

S'ha fet la instal·lació d'un códec de vídeo MPEG-4 perquè funcionés correctament la classe.

*NTGraph*

S'utilitza un control OCX que permet fer representacions gràfiques en dos dimensions. Aquest control és capaç de representar un alt nombre de punts i actualitzar les gràfiques amb noves dades.

### 6.1.2. Estructura

Per implementar el descriptor de textura s'han seguit les especificacions del descriptor de textura de l'estàndard MPEG-7 explicat en 4.1.

No obstant, hi ha algunes diferències ja que no s'ha implementat el sistema de visió humana HSV. Enlloc del sistema HSV s'ha utilitzat la transformada de Fourier en dos dimensions de la imatge.

El diagrama que s'ha seguit per a l'extracció de les característiques és el de la Fig. 6.2.1.

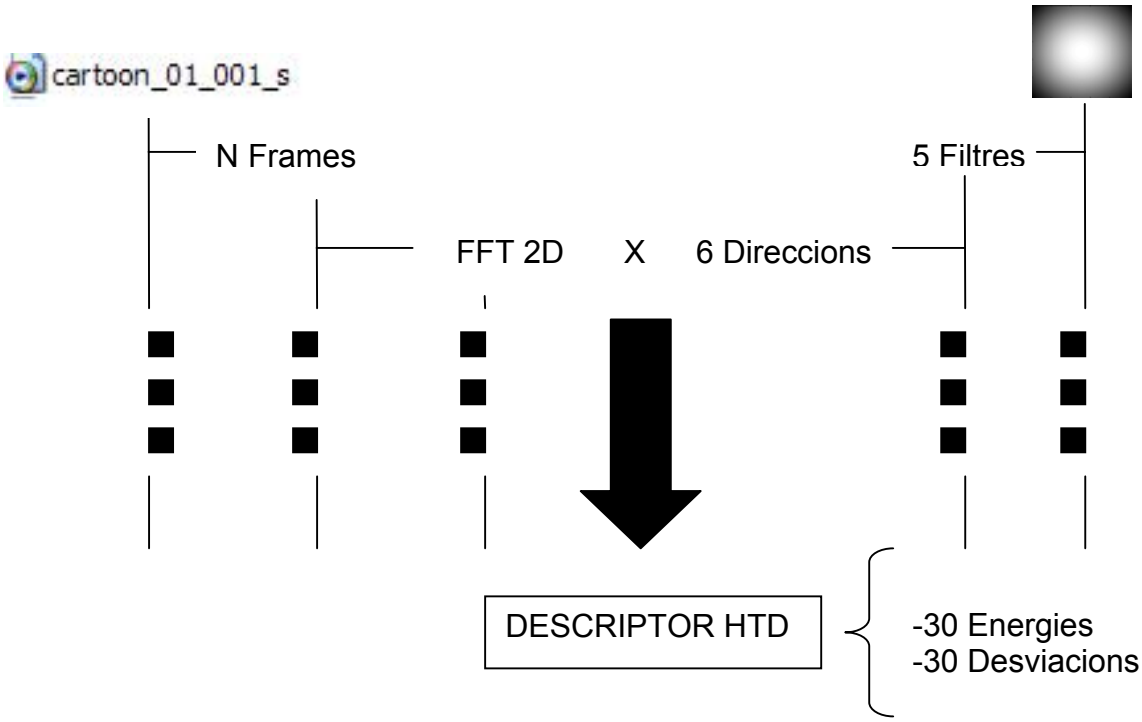



**Fig. 6.2.1** Estructura per l'extracció del HTD

A partir d'un vídeo s'han agafat els seus frames, i de cada frame s'ha realitzat la FFT 2D. El resultat de la FFT 2D s'ha normalitzat per tal que les freqüències baixes es trobin situades al centre de la imatge i les altes als extrems de la imatge. Per dur a terme aquestes operacions s'han utilitzat conjuntament les llibreries de Intel Ipl i Ippi, i les OpenCV.

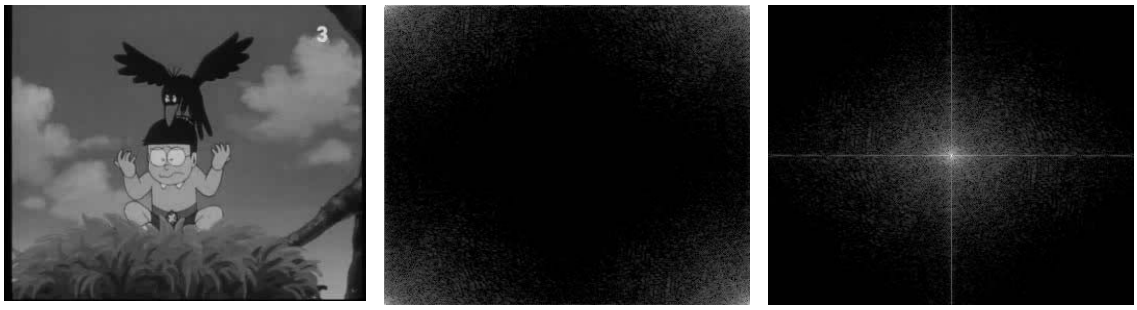


**Fig. 6.2.2** a) Frame Original b) FFT-2D sense normalitzar (baixa freqüència a les cantonades de la imatge), c) FFT-2D normalitzada

Totes les operacions realitzades s'han fet amb imatges d'un canal, i per efectuar operacions entre elles és recomanable que tinguin una profunditat de 32 bits perquè no es perdi informació.

Per altra banda, s'han creat els filtres Gabor. La dificultat d'aquests filtres 2D és que una variable és angular i per tant s'han de dur a terme interpolacions per situar els valors de la gaussiana a la direcció desitjada.

El filtre especificat en l'estàndard MPEG-7 està normalitzat i també les seves desviacions angulars i radials. En aquest projecte els valors de la Taula. 4.1.1 i Taula. 4.1.2 s´han calculat respecte les dimensions 352 x 288.

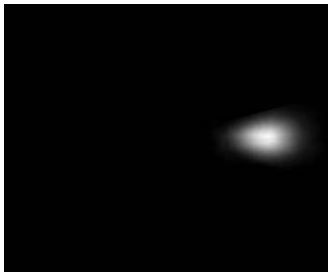


**Fig. 6.2.3** Filtre Gabor nº1

Una vegada obtinguts els filtres i les transformades FFT 2D es procedeix a la seva multiplicació per obtenir les energies de cada filtre i les desviacions com s'observa en la Fig. 6.2.4.

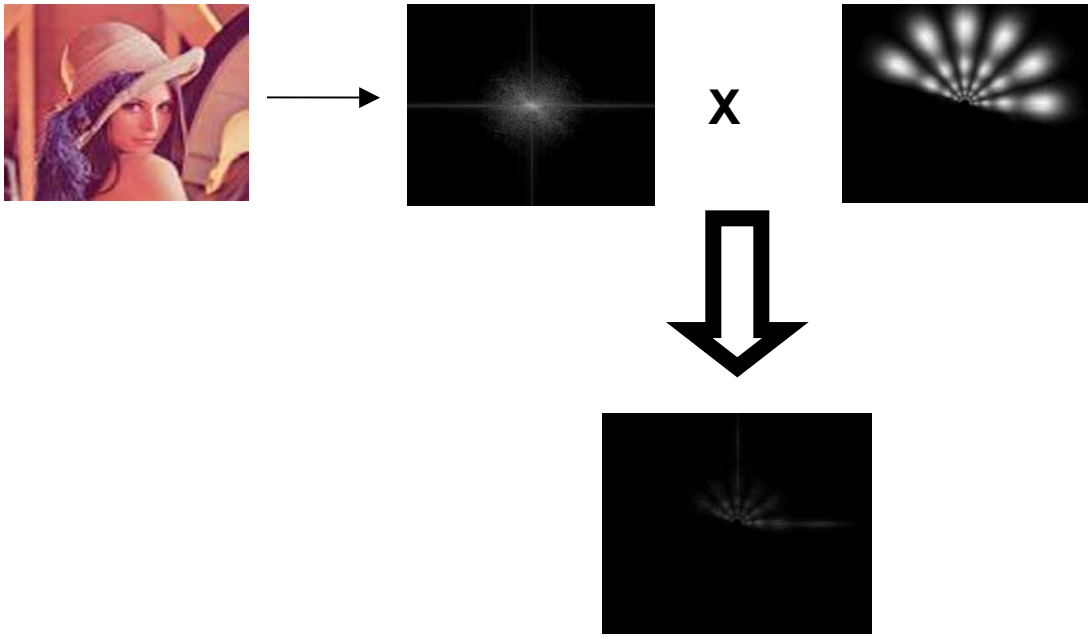



**Fig. 6.2.4** Obtenció de la potència i desviacions

Les classes que realitzen tot aquestes operacions s'agrupen en tres CFilterGabor, CVideoTexture i CVideoTextureDlg.

*CFilterGabor*

Aquesta classe és l'encarregada de generar els 30 filtres gabor. A partir de la funció de Gauss en 2D es van creant els filtres segons l'angle, la sigma_x, la sigma_y i la banda. El resultat final és l'obtenció de 30 imatges en format jpg que faran de màscara en la FFT 2D de cada frame. A continuació hi ha un diagrama que explica l'estructuració d'aquesta classe.

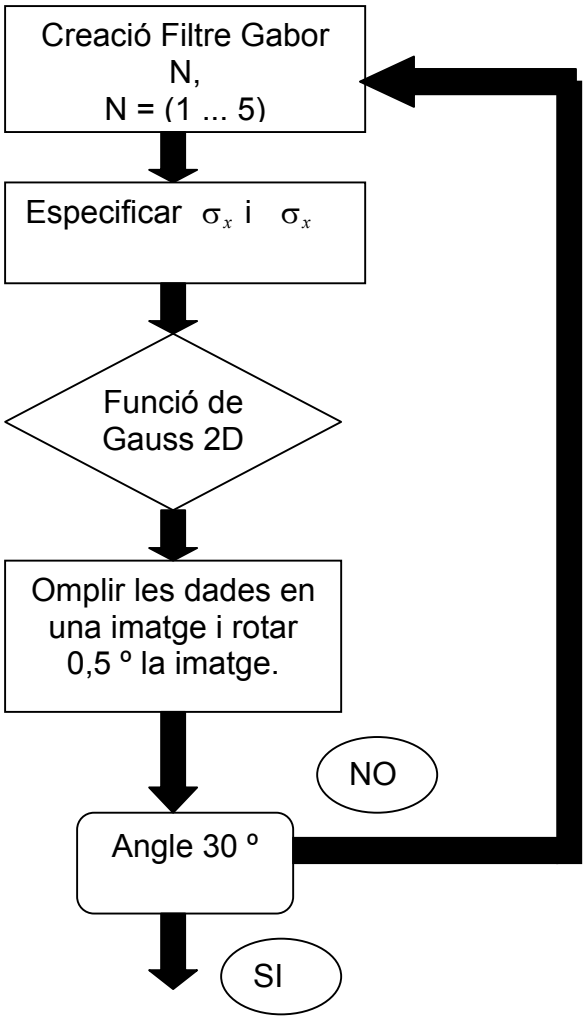



**Fig. 6.2.5** Diagrama de la classe CFilterGabor

Una vegada obtinguts els cinc filtres en la direcció 0º, només s'han d'anar rotant la imatge cada 30 graus per obtenir els filtres en les altres direccions.

*CVideoTexture*

Aquesta classe és l'encarregada d'obtenir el descriptor HTD. El diagrama de la classe es detalla en la Fig. 6.2.6.

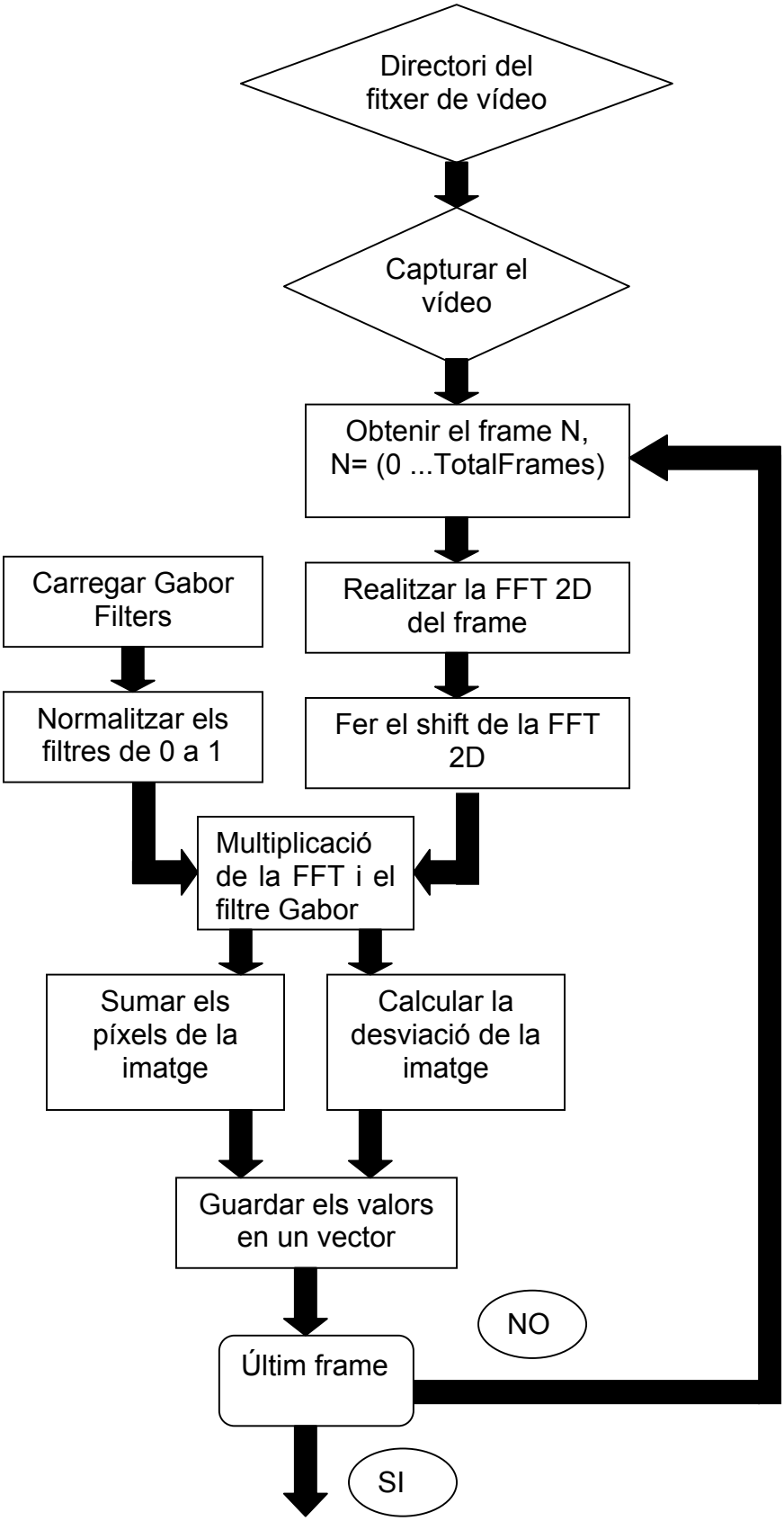



**Fig. 6.2.6** Diagrama de la classe CVideoTexture

*CVideoTextureDlg*

Aquesta classe utilitza les dues anteriors, i és una interfície gràfica per poder visualitzar la FFT 2D i la FFT 2D amb els filtres Gabor.

Per incloure aquestes tres classes en el projecte gobal explicat en l'apartat 1.2, s'han fet diverses modificacions a les altres classes disponibles.

*CAVFeaturesSelectionDlg*

Aquesta classe ja estava implementada anteriorment, però s'han hagut de fer modificacions per incloure el descriptor de textura. La finalitat d'aquesta classe és poder seleccionar totes les característiques desitjades per l'extracció d'aquestes en un fitxer de text.

*CVideoFeaturesExtractionDlg*

Aquesta classe consisteix en crear un fitxer de text amb les característiques calculades sobre els fitxers audiovisuals. Aquesta classe també estava implementada, i s'han fet modificacions per incloure el descriptor HTD.

### 6.1.3. Funcionament

S'ha realitzat una interfície gràfica per poder visualitzar aquest procés (*CVideoTextureDlg*). Com es veu en la Fig. 6.2.7 la part esquerra de la interfície hi ha el frame del vídeo i en la part dreta la FFT 2D d'aquest frame.

Les accions que et permet fer aquesta interfície és avançar o retrocedir la seqüència del vídeo, aplicar un filtre en la FFT 2D en la direcció desitjada, i accentuar la FFT 2D variant el valor de c, on c correspon al valor de la formula Imatge = c · log [FFT_Image].

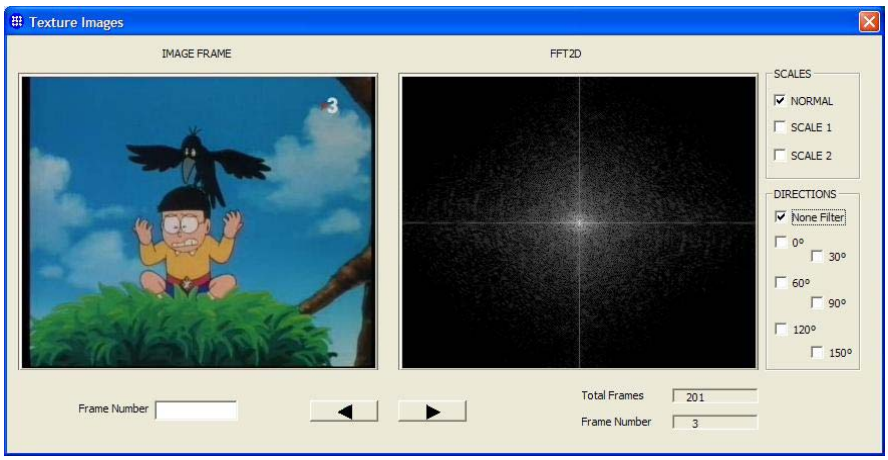



**Fig. 6.2.7** Interfície Gràfica

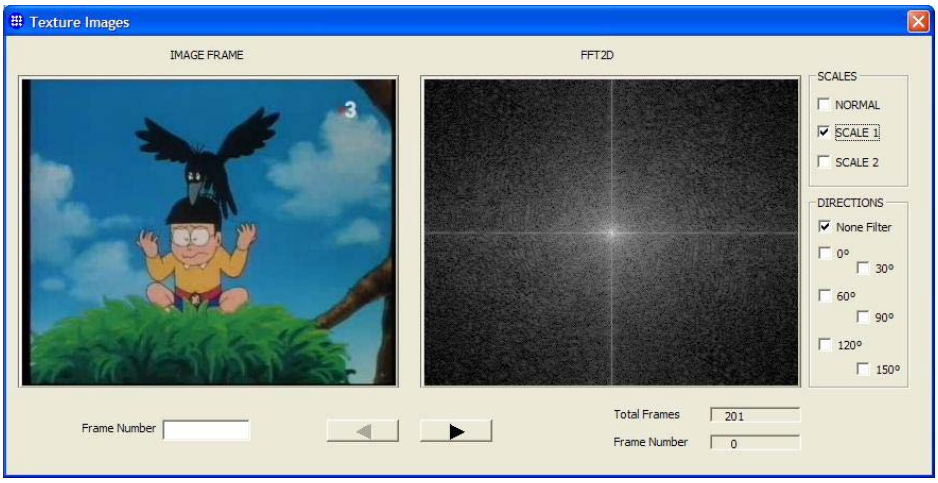



**Fig. 6.2.8** Modificar l'escala de la FFT 2D

Es pot observar en la Fig. 6.2.8 que quan modifiquem l'escala (el valor de c) la FFT 2D s'expandeix i els píxels on abans es veien de color negre es converteixen en tons més blancs (veure [7] ) .

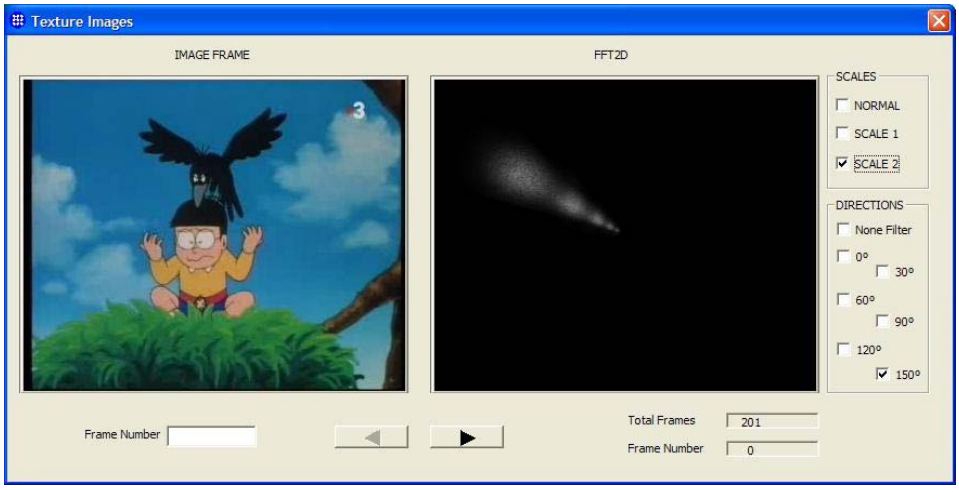



**Fig. 6.2.9** Filtre en la direcció 150º a la FFT 2D

En aquesta altra imatge es pot veure l'efecte d'aplicar els filtres Gabor nº 6,12, 18 ,24 i 30 que corresponen a la direcció 150 º.

En resum, s'han creat tres classes en Visual C++, s'ha unificat amb el programa ja existent d'extracció de característiques d'arxius multimèdia i s'ha creat un help per consultar l'estructura de classes i les explicacions en els arxius .h.

# CAPÍTOL 7. RESULTATS

En aquest capítol s'explicaran els resultats obtinguts en les proves que s'han fet en la classificació i el material generat durant el projecte.

## 7.1. Segments de vídeos

Al final d'aquest projecte es disposa d'una base de dades de dibuixos animats. S'han creat 80 segments a partir de dos vídeos de dibuixos animats gravats directament del TV3 de 40 minuts cadascun.

**Taula. 7.1.1** Fitxers creats per a cada segment

| Noms | Descripció |
|---|---|
| Cartoon_0X_XXX.mpg | DVD-Format (720x560) Video segments |
| Cartoon_0X_XXX.wav | Audio segments |
| Cartoon_0X_XXX_s.avi | CIF-Format (352x288) Video segments |
| Cartoon_0X_0XX.txt | Etiquetes per a cada segment |
| Cartoon_0X.txt | TXT amb totes les etiquetes anteriors |

En total s'ha aconseguit una base de dades formada per 140 fitxers (de text, d'àudio i vídeo).

Cal comentar que la segmentació del vídeo en escenes és el primer pas per fer qualsevol classificació.

## 7.2. Extraccions de característiques

Per l'obtenció d'un fitxer de característiques, inicialment només s'ha fet l'anàlisi de contingut audiovisual amb el descriptor de textura, en el nostre cas el HTD.

Segons el procés explicat en el 6.2, s'han calculat els 62 valors del descriptor de textura per cada frame i s'han guardat en una matriu (energies i desviacions). Posteriorment s'ha creat un fitxer de text amb tots aquests valors.

Aquest fitxer està format per 63 columnes i N+2 files, on N és el nombre de frames del vídeo. També s'han generat fitxers de text més reduïts agrupant la informació en menys característiques.

El procés d'extracció de característiques és lent ja que s'han d'extreure les característiques de tots els segments de vídeos disponibles, tan de dibuixos animats, com de no-dibuixos animats.

El resultat final són 80 fitxers de texts referents a dibuixos animats i 138 fitxers de texts referents a no-dibuixos animats.

VARIS_01_001_s_VideoFeatures - Bloc de notas

Archivo Edición Formato Ver Ayuda

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | FRAME LEVEL FEATURES | | |
| TC | MeanHTD | DsvHTD | GabFil01 | GabFil02 | GabFil03 | GabFil04 |
| 0 | 86.15823 | 27.12377 | 11.31664 | 22.78895 | 11.37827 | 5.016043 |
| 1920 | 86.32405 | 27.23102 | 12.07200 | 22.47154 | 10.93536 | 4.799129 |
| 3840 | 86.79785 | 27.24775 | 11.97640 | 22.57902 | 10.94025 | 4.795575 |
| 5760 | 86.21712 | 27.02850 | 12.33899 | 22.33558 | 10.79639 | 4.729583 |
| 7680 | 86.93952 | 26.86626 | 12.05991 | 22.50712 | 10.93670 | 4.776406 |
| 9600 | 86.96918 | 26.73026 | 12.00625 | 22.56823 | 10.92366 | 4.779804 |
| 11520 | 86.39900 | 26.58589 | 12.40393 | 22.36335 | 10.71942 | 4.665646 |
| 13440 | 87.06402 | 26.54801 | 12.22339 | 22.41755 | 10.87043 | 4.710550 |
| 15360 | 87.23760 | 26.46697 | 11.83565 | 22.63631 | 11.04099 | 4.812488 |
| 17280 | 86.80703 | 26.40676 | 12.41597 | 22.27458 | 10.78679 | 4.702940 |
| 19200 | 87.21281 | 26.53099 | 12.42326 | 22.24226 | 10.76343 | 4.719771 |
| 21120 | 87.21377 | 26.52984 | 12.46868 | 22.22637 | 10.74618 | 4.716834 |
| 23040 | 87.22991 | 26.53561 | 12.52134 | 22.22238 | 10.72149 | 4.700500 |
| 24960 | 87.10256 | 26.54758 | 12.71660 | 22.02459 | 10.67495 | 4.681794 |
| 26880 | 86.32409 | 26.56677 | 12.79201 | 21.89372 | 10.69742 | 4.701526 |
| 28800 | 86.58054 | 26.68279 | 12.84494 | 21.93476 | 10.62166 | 4.673318 |
| 30720 | 86.56370 | 26.88681 | 12.41140 | 22.14221 | 10.81667 | 4.798072 |
| 32640 | 86.22305 | 27.09470 | 12.81381 | 21.85939 | 10.72397 | 4.717898 |
| 34560 | 86.80419 | 27.30716 | 12.69620 | 21.99273 | 10.73537 | 4.747428 |
| 36480 | 87.10714 | 27.46040 | 12.49980 | 22.04632 | 10.86831 | 4.792679 |
| 38400 | 86.88936 | 27.52187 | 12.72407 | 21.90181 | 10.81510 | 4.750917 |
| 40320 | 87.52732 | 27.58976 | 12.25346 | 22.20816 | 11.00882 | 4.827019 |
| 42240 | 87.57748 | 27.61797 | 12.23826 | 22.21473 | 11.02523 | 4.836084 |
| 44160 | 87.43720 | 27.60372 | 12.52187 | 22.08932 | 10.85503 | 4.786975 |
| 46080 | 87.72770 | 27.55664 | 12.57268 | 22.04369 | 10.85362 | 4.758429 |
| 48000 | 87.62881 | 27.52597 | 12.81442 | 21.87865 | 10.72395 | 4.709152 |
| 49920 | 87.28131 | 27.38198 | 12.87072 | 21.89426 | 10.67932 | 4.678469 |
| 51840 | 87.34482 | 27.35373 | 13.22073 | 21.73485 | 10.45198 | 4.610823 |
| 53760 | 87.21411 | 27.26493 | 12.76504 | 22.01904 | 10.64674 | 4.703954 |
| 55680 | 86.26686 | 27.05044 | 12.80648 | 21.97683 | 10.63178 | 4.701078 |

Línea 1, columna 1

**Fig. 7.2.1** Fitxer de text amb les característiques de l'HTD

Finalment, amb el programa MacroFiles (veure 6.1) s'han unit els fitxers associats a cada segment segons si són de dibuixos animats o no-dibuixos animats, reduint-se els fitxers en només quatre: 2 macrofitxers d'entrenament i 2 marofitxers de test. Aquests 4 fitxers són els que s'utilitzaran per avaluar els classificadors. Els fitxers s'han anomenat CARTRAIN.txt (40 fitxers associats), NOCARTRAIN.txt (69 fitxers associats), CARTEST.txt (40 fitxers associats) i NOCARTEST.txt (69 fitxers associats).

Per dur a terme una classificació multimodal s'ha seguit el mateix procediment explicat en tot aquest apartat, però s'han extret fitxers de característiques d'àudio i vídeo conjuntament. La generació d'aquests fitxers necessita un alt càlcul computacional. En cada fitxer de característiques hi han 122 característiques, 25 d'àudio a nivell de frame, 18 d'àudio a nivell de clip i 79 de vídeo.

A partir d'aquests fitxers de 122 característiques es poden fer multitud de proves durant la classificació.

## 7.3. Classificació de dibuixos animats

A partir dels fitxers d'entrenament (NOCARTRAIN.txt i CARTRAIN.txt) i de test (NOCARTEST.txt i CARTEST.txt) s'han fet diverses proves de classificació.

S'han fet els entrenaments i els tests a través dels classificadors explicats en el capítol 5.

La classificació s'ha fet a través d'uns programes realitzats en Matlab. Per a cada classificador s'ha disposat de tres arxius .m.

El primer arxiu anomenat ReadAVFeatures.m agafa la informació del macrofitxer i col·loca les dades en una matriu transposada.

L'altre arxiu és l'encarregat de l'entrenament del classificador i s'anomena XXXTraining.m .

I l'últim s'encarrega de fer el testeig de la classificació, anomenat XXXTest.m.

Per facilitar la tasca a l'hora d'executar els tres arxius s'ha creat un altre arxiu .m on els *inputs* són els quatre noms dels macrofitxers, i els *outputs* la taxa de reconeixement dels segments de vídeo segons si són dibuixos animats o no són dibuixos animats.

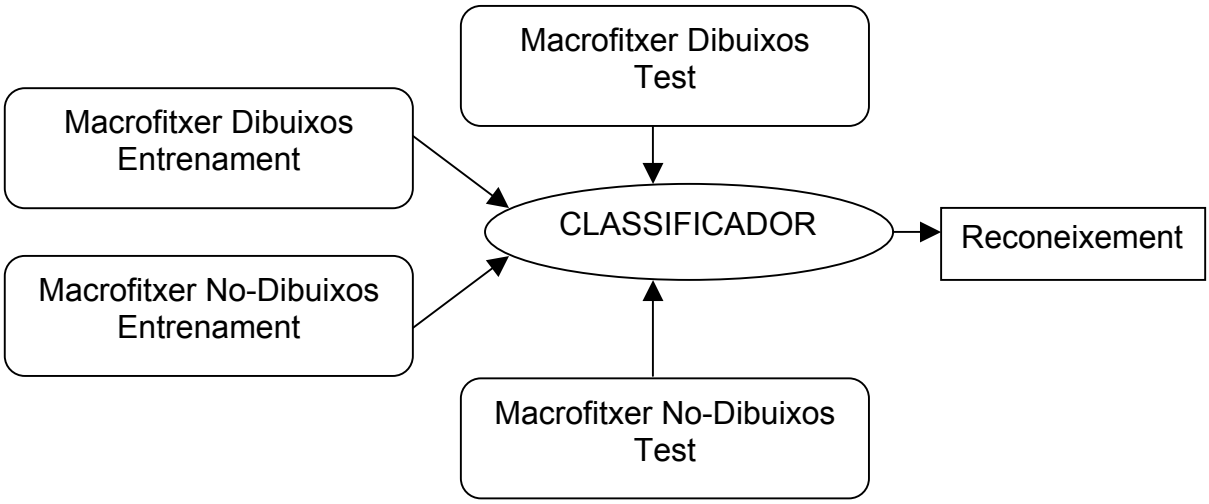



**Fig. 7.3.1** *Inputs* i *Outputs* en la classificació

### 7.3.1. Classificació amb el descriptor de textura

Inicialment, s'ha fet un entrenament utilitzant només el descriptor bàsic HTD format per les 30 potències i les fdc (desviació) i fsc (mitja).

En la Taula. 7.3.1 es mostren aquestes proves on les columnes de la taula hi ha els classificadors utilitzats (PCA, LDA i M-NN), i en les files les diverses proves que s'han efectuat.

Per tal que les característiques dels macrofitxers d'entrenament tinguin el mateix rang i donar a les característiques la mateixa importància, s'han normalitzat els seus valors numèrics (valor numèric de 0 a 1) abans de fer l'entrenament i el test.

En aquest apartat no seria necessari normalitzar les característiques ja que només s'utilitza el descriptor de textura, no obstant també s'han fet proves normalitzant els valors del descriptor.

És per això que en la taula apareixen dos divisions per a cada classificador, el primer cas (NORM) les característiques estan normalitzades, i en el segon cas no (NO-NORM).

Els valors numèrics que hi ha a la taula són les taxes de reconeixement en tant per u. És a dir, que si el valor numèric és 0,7 ha reconegut el 70% dels segments de vídeo de dibuixos animats.

**Taula. 7.3.1** Resultats inicials

| | **PCA** | | **LDA** | | **M-NN** | |
|---|---|---|---|---|---|---|
| | NORM | NO-NORM | NORM | NO-NORM | NORM | NO-NORM |
| HTD [32] | 0,3 | 0,7 | 0,5 | 0,5 | 0,6 | 0,7 |
| 0º | 0,713 | 0,78 | 0,6 | 0,52 | 0,38 | 0,47 |
| 30º | 0,69 | 0,76 | 0,404 | 0,404 | 0,6 | 0,71 |
| 60º | 0,3 | 0,33 | 0,55 | 0,59 | 0,64 | 0,53 |
| 90º | 0,66 | 0,75 | 0,77 | 0,77 | 0,72 | 0,75 |
| 120º | 0,372 | 0,3 | 0,52 | 0,52 | 0,63 | 0,5 |
| 150º | 0,332 | 0,7 | 0,25 | 0,25 | 0,49 | 0,27 |
| Banda 1 | 0,34 | 0,82 | 0,51 | 0,51 | 0,3 | 0,35 |
| Banda 2 | 0,36 | 0,29 | 0,45 | 0,45 | 0,3 | 0,31 |
| Banda 3 | 0,34 | 0,34 | 0,64 | 0,64 | 0,34 | 0,34 |
| Banda 4 | 0,67 | 0,64 | 0,56 | 0,56 | 0,64 | 0,64 |
| Banda 5 | 0,73 | 0,64 | 0,52 | 0,52 | 0,71 | 0,65 |

Tal com s'observa en la Taula. 7.3.1, la primera prova (HTD[32]) s'ha fet una classificació utilitzant els 32 valors de la HTD bàsica (30 energies, desviació (fdc) i la mitja (fsc)).

Els resultats no han estat brillants en el classificador LDA, però en els altres s'ha arribat a un reconeixement del 70 %.

Després, com es pot veure en la Taula. 7.3.1 s'han anat agafant menys valors per observar el reconeixement segons la direcció (0º,30º, 60º, 90º, 120º i 150º) i la banda (Banda 1, Banda 2, etc.). S'ha vist que en la direcció 90º és on hi ha el millor reconeixement i en les bandes pròximes a baixa freqüència. Això és degut a què la informació freqüencial de les imatges és majoritàriament a baixes freqüències.

### 7.3.2. Classificació amb el descriptor de textura reduït

Un altre prova que s'ha fet és sumar les energies de cada banda i fer la mitja, aconseguint reduir fins a 12 el nombre de característiques. També s'ha inclòs el classificador Fuzzy Integral en la taula de resultats. En aquest cas els resultats obtinguts no han estat els esperats.

El què es volia fer amb aquesta prova és fer una anàlisi a més alt nivell, fent la mitja de les energies de cada banda sense tenir en compte les direccions.

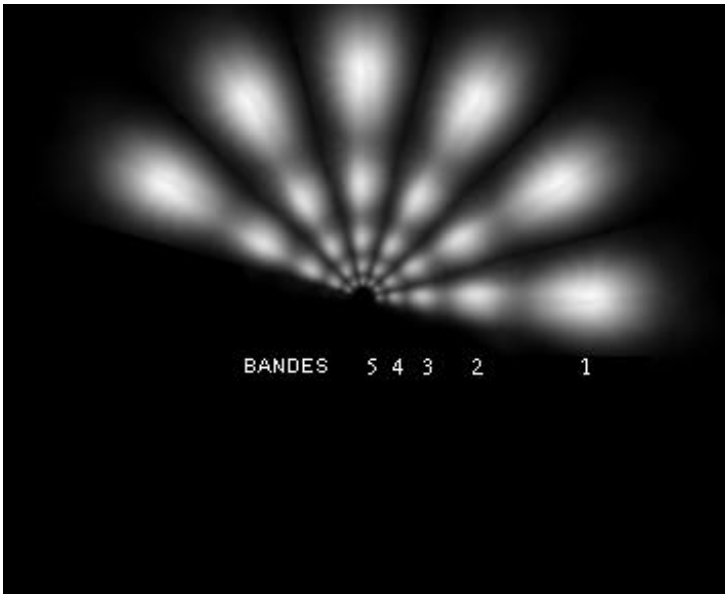



**Fig. 7.3.2** Bandes del filtre Gabor

A la Taula. 7.3.2 s'observen els resultats de les taxes de reconeixement on la màxima taxa arriba al 58%.

Això significa que realment no s'aprecia una diferència notable entre les bandes en vídeos de dibuixos i les bandes en vídeos de no-dibuixos, i que realment la diferència més important està en les direccions.

**Taula. 7.3.2** Classificació amb descriptor de textura reduït

| | **PCA** | | **LDA** | | **M-NN** | | **FUZZY** |
|---|---|---|---|---|---|---|---|
| | NORM | NO-NORM | NORM | NO-NORM | NORM | NO-NORM | NORM |
| HTDv2 | 0,53 | 0,58 | 0,4 | 0,42 | 0,36 | 0,53 | 0,54 |

### 7.3.3. Classificació multimodal

A continuació, s'ha procedit a fer una classificació utilitzant característiques d'àudio i vídeo conjuntament. En la Taula. 7.3.3 es mostren les tres proves que s'han realitzat. S'han escollit característiques d'àudio a nivell de frame (AF), característiques d'àudio a nivell de clip (AC) i característiques de vídeo (AV).

L'elecció de les característiques més rellevants ha estat basant-se en dos estudis de classificació, un sobre la classificació d'àudio en la parla (veure [11] ), i l'altre en la classificació de vídeo (veure [12] ).

**Taula. 7.3.3** Classificació amb anàlisi multimodal

| | | | PCA | | LDA | | M-NN | | FUZZY |
|---|---|---|---|---|---|---|---|---|---|
| AF | AC | V | NORM | NO-NORM | NORM | NO-NORM | NORM | NO-NORM | NORM |
| 25 | 18 | 79 | 0,71 | 0,69 | - | - | 0,67 | 0,59 | - |
| 4 | 3 | 11 | 0,75 | 0,52 | 0,83 | 0,83 | 0,68 | 0,48 | 0,8 |
| 2 | 3 | 7 | 0,72 | 0,69 | 0,81 | 0,81 | 0,8 | 0,7 | 0,78 |

Primer s'ha fet una classificació utilitzant totes les característiques disponibles (122 característiques).

En la segona prova s'han agafat 4 característiques d'àudio a nivell de frame (Pitch, Centroid, Bandwidth i Rollof), 3 característiques d'àudio a nivell de clip (Mean-Vol, FC-Mean i BW-Mean) i 11 característiques de vídeo (Mean-HSV, Var-HSV, EntrHSV, SkeHSV, KurtoHSV, MeanPCF, VarPCF, Gabor1, Gabor25, HTD_Desv1 i HTD_Desv25). Els resultats han millorat notablement, sobretot en el classificador LDA, que ha aconseguit un reconeixement del 83%.

En la tercera prova s'han agafat 2 característiques d'àudio a nivell de frame (Pitch i Rolloff ), 3 característiques d'àudio a nivell de clip (FC-Mean,BW-Mean i ESBR1-Me) i 7 característiques de vídeo (EntrpHSV, Autocorr, KurtoPCF, Gabor1, Gabor25, HTD_Desv1 i HTD_Desv25). Les taxes de reconeixement han estat similars a les anteriors, menys en el classificador M-NN que ha assolit el 81%.

### 7.3.4. Classificació multimodal amb finestra temporal

Després, s'ha provat d'utilitzar una finestra temporal d'un segon per avaluar un nivell més alt que el nivell de frame, i veure si el reconeixement millorava. Com es pot observar en la Fig. 7.3.3, la finestra temporal utilitzada té un solapament amb l'anterior.

**Fig. 7.3.3** Finestres temporals solapades

S'ha fet la mediana de cada finestra de 25 frames obtenint un únic valor numèric per a cada característica.

Els resultats de reconeixement han millorat respecte les proves realitzades sense utilitzar finestres temporals.

**Taula. 7.3.4** Classificació utilitzant finestres temporals en els macrofitxers

| | | | **PCA** | **LDA** | **M-NN** | **FUZZY** |
|---|---|---|---|---|---|---|
| AF | AC | V | NORM | NORM | NORM | NORM |
| 4 | 3 | 11 | 0,81 | 0,85 | 0,69 | 0,78 |
| 2 | 3 | 7 | 0,77 | 0,81 | 0,81 | 0,75 |

Utilitzant les finestres temporals i fent la mediana s'obtenen uns valors de les característiques més generals que en aquest cas afavoreix a la classificació del gènere de dibuixos animats.

# CAPÍTOL 8. CONCLUSIONS I LÍNIES FUTURES

La temàtica d'aquest projecte (l'anàlisi de contingut audiovisual) és d'actualitat, i crec que el seu interès anirà creixent amb l'augment de continguts audiovisuals i nous canals de distribució de continguts.

El descriptor de textura utilitzat per a la classificació del gènere de dibuixos animats ha aconseguit un reconeixement del 70%. Tenint en compte que només s'ha utilitzat un descriptor és un resultat bastant bo. Per tant, es pot dir que la seva elecció per la classificació de dibuixos animats ha estat correcte.

Aquest descriptor de textura està definit a l'estàndard MPEG-7. És important remarcar que l'estàndard MPEG-7 està enfocat a la indexació del contingut audiovisual, i deixa més oberta la part de classificació. És a dir, defineix una sèrie de descriptors, però no defineix cap tipus de classificador, ni especifica la importància de cada descriptor pels diferents tipus de contingut audiovisual.

Referent a la classificació multimodal (característiques d'àudio i vídeo), s'ha pogut observar que ofereix millors resultats. L'àudio aporta informació addicional que no aporta el vídeo. S'ha aconseguit un reconeixement del 85% en el millor cas, i els resultats del reconeixement han estat similars als obtinguts per Glasberg [9] .

Les llibreries d'Intel utilitzades en el projecte han estat de gran utilitat per realitzar totes les operacions de processat d'imatge. No han presentat problemes de compatibilitat entre elles, i s'han complementat molt bé. La llibreria IPP té funcionalitats més específiques, la IPL és més genèrica i la OpenCV queda entre les dues.

La classificació del gènere dibuixos animats a temps real encara queda lluny, ja que es necessita un cert temps d'anàlisi per determinar el gènere. Encara que s'hagi fet una segmentació del vídeo en escenes per facilitar l'extracció de característiques, la seva extracció necessita més temps del desitjable.

Fent el projecte han anat apareixent moltes idees que s'han quedat en la ment, i no s'han pogut portar a la pràctica. A continuació n'esmento algunes.

La taxa de reconeixement de dibuixos animats podria millorar si s'afegeixen més descriptors, però també s'ha de vigilar amb l'elecció dels descriptors utilitzats en cada cas perquè pot ser que no siguin representatius. Aquest és un problema que podria solucionar el classificador Fuzzy Integral, ja que podria determinar els descriptors més representatius segons la classificació.

Per aconseguir tots els resultats abans s'han hagut de seguir una sèrie de passos (segmentació manual, anàlisi dels segments, extracció de característiques etc.) que seria interessant automatitzar per agilitzar el procés. Per exemple la segmentació automàtica dels fitxers de vídeo en escenes.

Un altre aspecte és el temps de processament de l'ordinador per extreure les característiques. Es podria fer l'anàlisi cada N frames per donar rapidesa al procés o utilitzar un ordinador més potent.

Respecte el descriptor de textura HTD, es podria implementar la transformada Radon abans d'obtenir la FFT 2D i observar si els resultats de la classificació milloren.

Complint la normativa de projectes de la universitat sobre l'impacte ambiental, comentar que aquest projecte no té cap efecte directe contra el medi ambient, i per tant no s'han pres mesures sobre el medi ambient.

# BIBLIOGRAFIA

[1] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Language,* 1st edition, 2002

[2] Yao Wang, Zhu Liu, and Jin-Cheng Huang, "Multimedia Content Analysis Using both Audio and Video Clues", *IEEE Signal Processing Magazine*, IEEE Inc., New York, NY, pp. 12 - 36, vol. 17, No 6, November 2000

[3] S. Fischer, R. Lienhart and W. Effelsberg, "Automatic recognition of film generes", in Proc. 3rd ACM International Conf. Multimedia, San Francisco, November 1995

[4] Y. Wang, Z. Liu and J.C Huang, "Multimedia content analysis using both audio and visual clues", *IEEE Signal Processing Magazine*, 2001

[5] O. D. Faugeras and W. K. Partt, "Decorrelation methods of texture feature extraction," IEE Trans. Pattern Anal. and Machine Intell., Vol.2, July 1980, pp. 323-332

[6] Gurewich O. Gurewich N, *Aprendiendo Visual C++ en 21 dias*, Prentice Hall ,1998

[7] Rafel C.,Richard E., *Digital Image Processing*, Addison-Wesley, September 1993

[8] Anil K.Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, 1989

[9] R. Glasberg, K. Elazouzi and T. Sikora*, Cartoon-Recognition using Visual-Descriptors and a Multilayer-Perceptron*, IAMIS, Montreux, April 13-15, 2005

[10] Grabisch. M, "Fuzzy Measures and Integrals for Decision Making and Pattern Recognition. Fuzzy Structures: Current Trends", Tatra Mountains Mathematical Publications, 1997.

[11] Juan D., "Sistema d'identificació a partir de l'anàlisi de la parla", Projecte Final de Carrera, EPSC, 29-06-2004.

[12] Ferré Ll., "Extracció de característiques i classificació de vídeo", Projecte Final de Carrera, EPSC, 29-06-2005.

[13] Forum about IPP Software Development Products
< http://softwareforums.intel.com>

[14] Software and samples about OpenCV
<http://sourceforge.net/projects/opencvlibrary/>

[15] Radon transform
<http://www.owlnet.rice.edu/~elec431/projects96/DSP/bpanalysis.html>

# ANNEXOS

## 1.1. Descripció de l'aplicació d'extracció de característiques

L'eina d'extracció de característiques està dividida en els blocs de la Fig. 1.1.1.

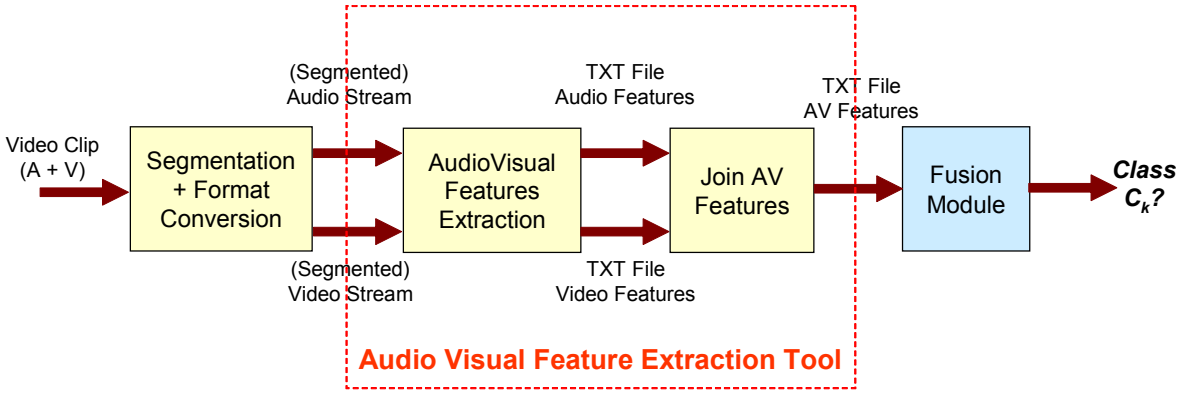



**Fig. 1.1.1** Diagrama de l'eina d'extracció de característiques

Per tal de simplificar i automatitzar el procés d'entrenament, s'ha segmentat el contingut audiovisual en escenes. Cada escena consta d'un fitxer d'àudio i vídeo de la mateixa durada. Aquesta segmentació s'ha fet manualment a través de l'Adobe Premiere 6.5.

El format en el fitxer d'àudio és el wav, i el format en fitxer de vídeo és el MPEG-4. Veure els detalls de la Fig. 1.1.2 i la Fig. 1.1.3.

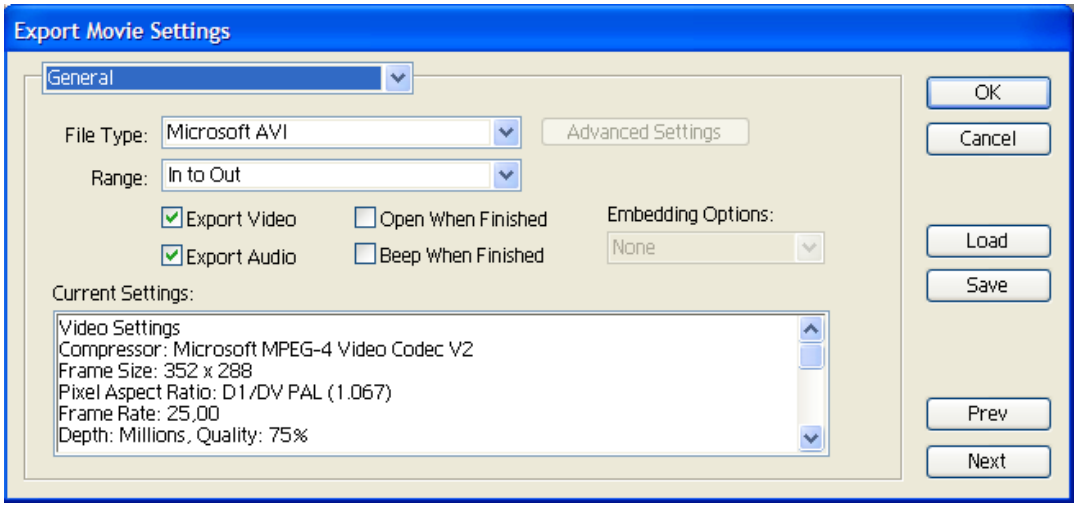



**Fig. 1.1.2** Format recomanat dels fitxers de vídeo utilitzant Adobe Premier 6.5

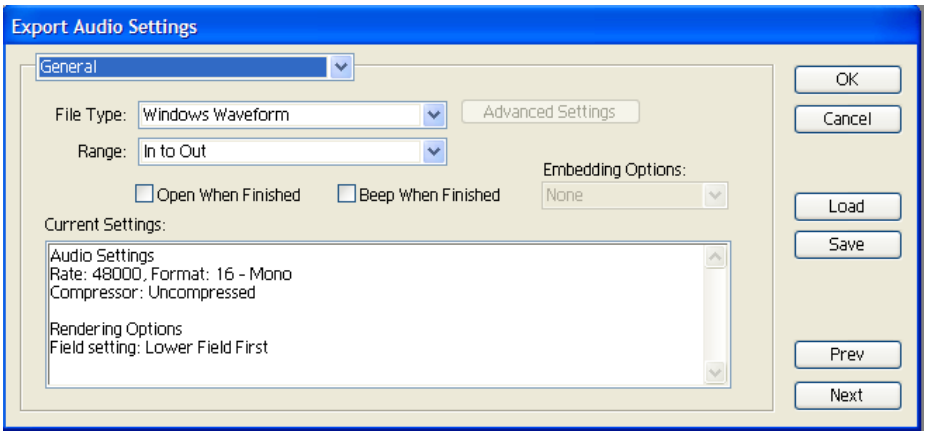



**Fig. 1.1.3** Format recomanat dels fitxers d'àudio utilitzant Adobe Premier 6.5

Una vegada l'àudio i el vídeo ha estat segmentat, els dos fitxers creats seran utilitzats com a paràmetres d'entrada en l'eina d'extracció de característiques. Primer s'especifiquen els directoris on s'han guardat els fitxers de vídeo i àudio (seleccionar l'etiqueta *same path* si tots dos, àudio i vídeo, són al mateix directori). Una altra opció és que pots seleccionar si vols processar només un fitxer d'àudio i vídeo, o processar tots els fitxers de vídeo i àudio del directori (seleccionar l'etiqueta *batch processing*). En la Fig. 1.1.4 es presenta la finestra principal del programa d'extracció de característiques.

Per processar tots els fitxers del directori, tots els fitxers han d'estar etiquetats amb el mateix nom "*Nom_XXX_s*" on *XXX* és l'identificador de cada fitxer. Per exemple, per processar un directori amb deu fitxers diferents, aquests han d'anomenar-se com *"Nom_001_s"* fins *"Nom_010_s"*, i marcar l'etiqueta *batch processing* en la finestra mostrada a la Fig. 1.1.4 .

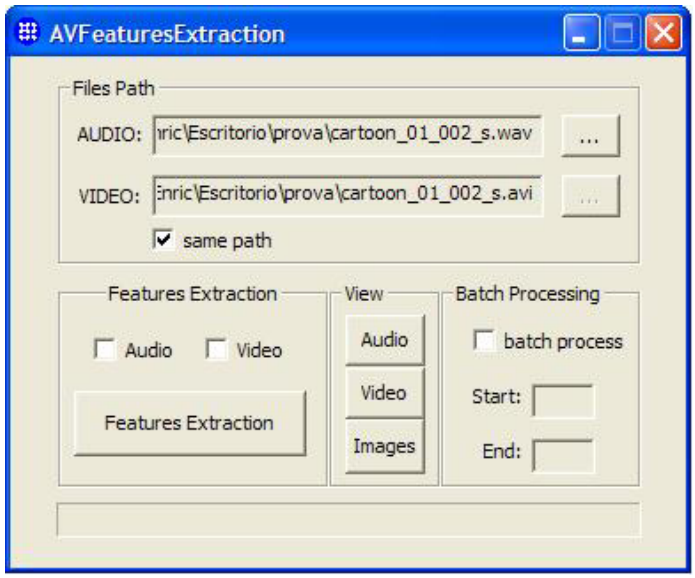



**Fig. 1.1.4** Finestra principal del programa d'extracció de característiques

L'eina també ofereix la possibilitat de veure i escoltar els fitxers de vídeo i àudio, i analitzar i representar les característiques separadament. Prement els Audio i Video l'aplicació mostra ambdues característiques mostrades a la figura Fig. 1.1.5 i a la Fig. 1.1.6.

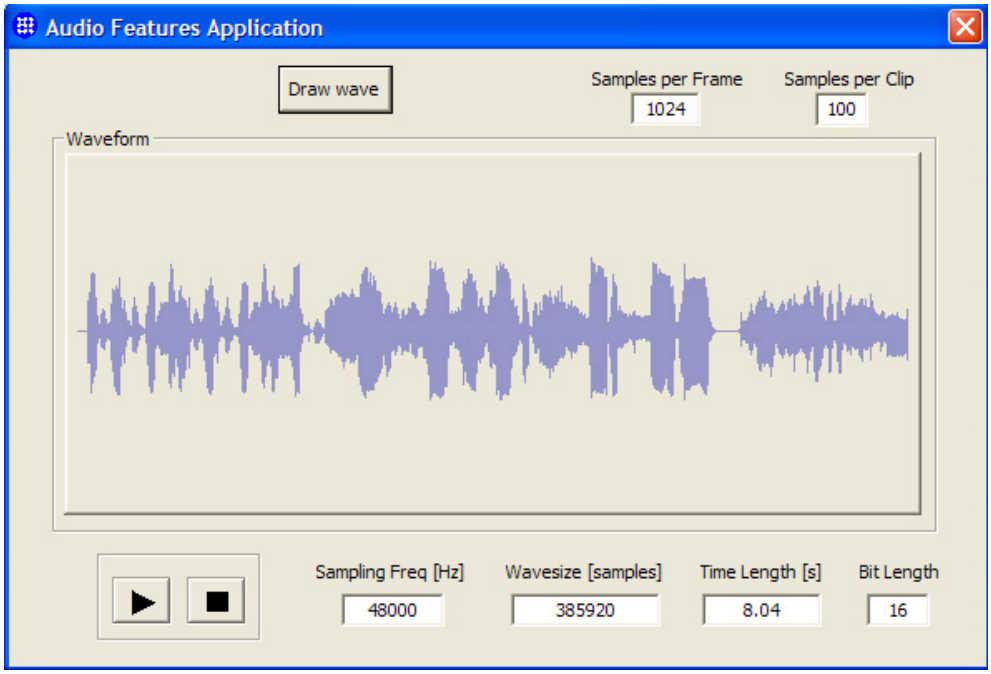



**Fig. 1.1.5** Visor de característiques d'àudio

Bàsicament, l'eina pot reproduir l'àudio, dibuixar la forma d'ona i obtenir alguna informació sobre el format del fitxer, com per exemple la freqüència de mostreig, la durada o els bits per mostra.

El visor de vídeo pot representar totes les característiques de vídeo seleccionades, i mostrar l'evolució d'aquestes durant el vídeo.

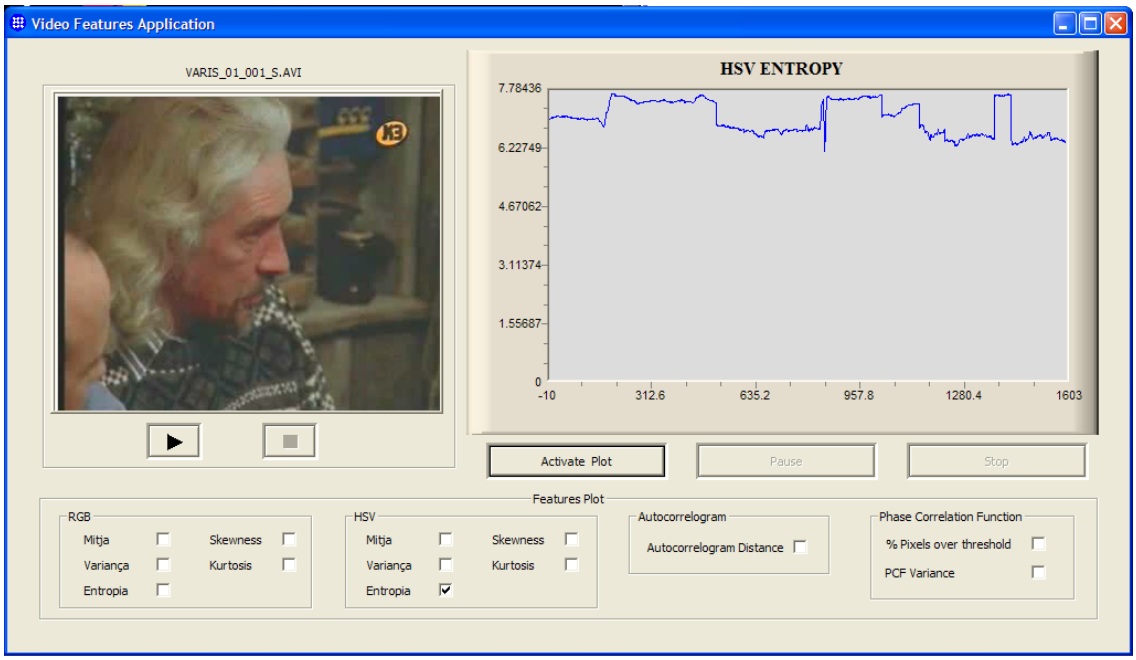



**Fig. 1.1.6** Visor de característiques de vídeo

Pulsant el botó *Image* de la finestra principal s'obra una finestra auxiliar que et permet veure els frames del fitxer de vídeo. També et permet veure la transformada FFT 2D, tal com s'observa en la Fig. 1.1.7.

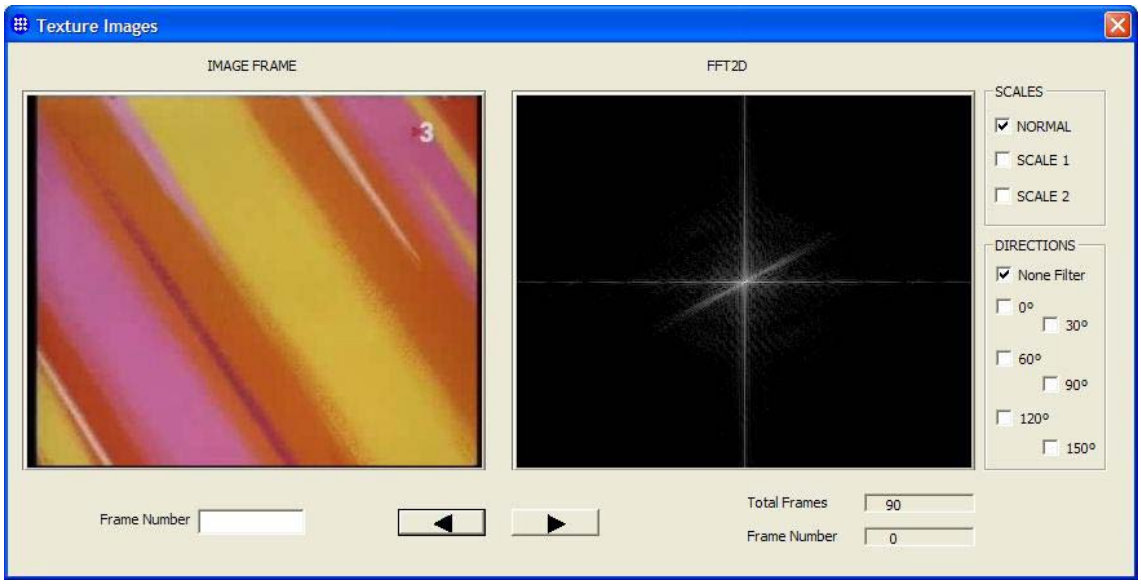



**Fig. 1.1.7** Visor de les FFT 2D

Aquest visor et permet també aplicar una màscara a la FFT 2D utilitzant els filtres Gabor. Pulsant les etiquetes "0º,30º,60º,90º," es pot veure diverses direccions de la FFT 2D.

En la finestra principal, prémer *Features Extraction Button* per seleccionar les característiques que es desitja analitzar i guardar en dos documents de text.

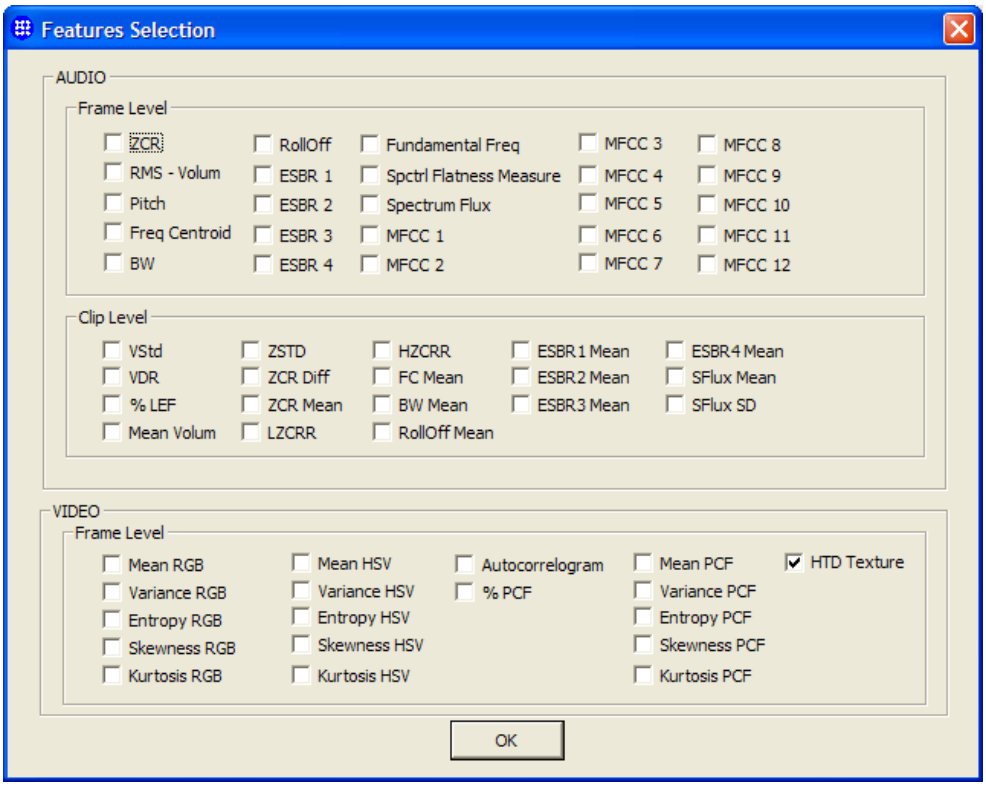



**Fig. 1.1.8** Selecció de característiques

L'estructura dels fitxers de text de sortida està representada a Fig. 1.1.9. Aquests dos fitxers poden ser utilitzats com a entrada d'un mòdul de classificació.

Totes les característiques d'àudio són analitzades utilitzant una freqüència de mostreig de 48 kHz i són classificats en dues categories diferents (audio frame level i audio clip level).

Les característiques de vídeo han estat analitzades només a nivell de frame.

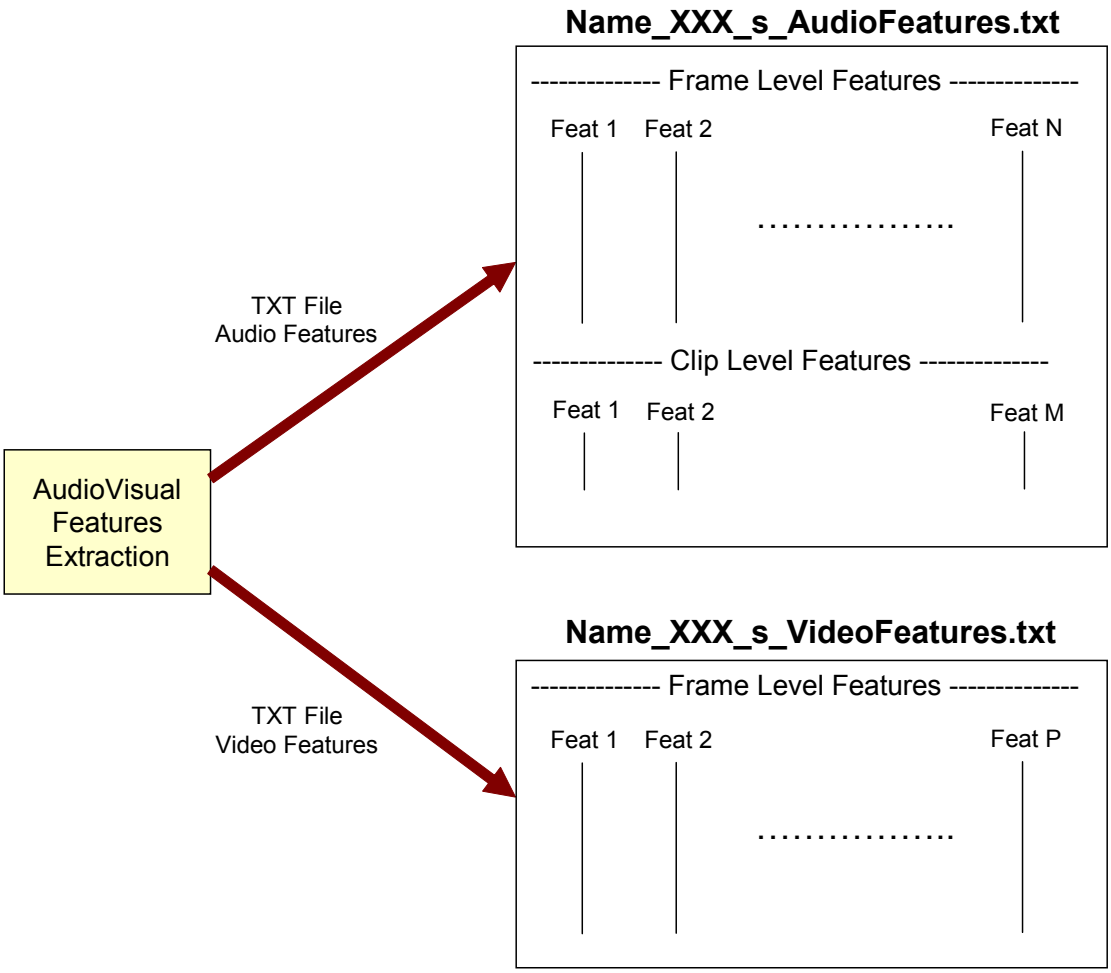



**Fig. 1.1.9** Estructura dels fitxers de sortida del programa

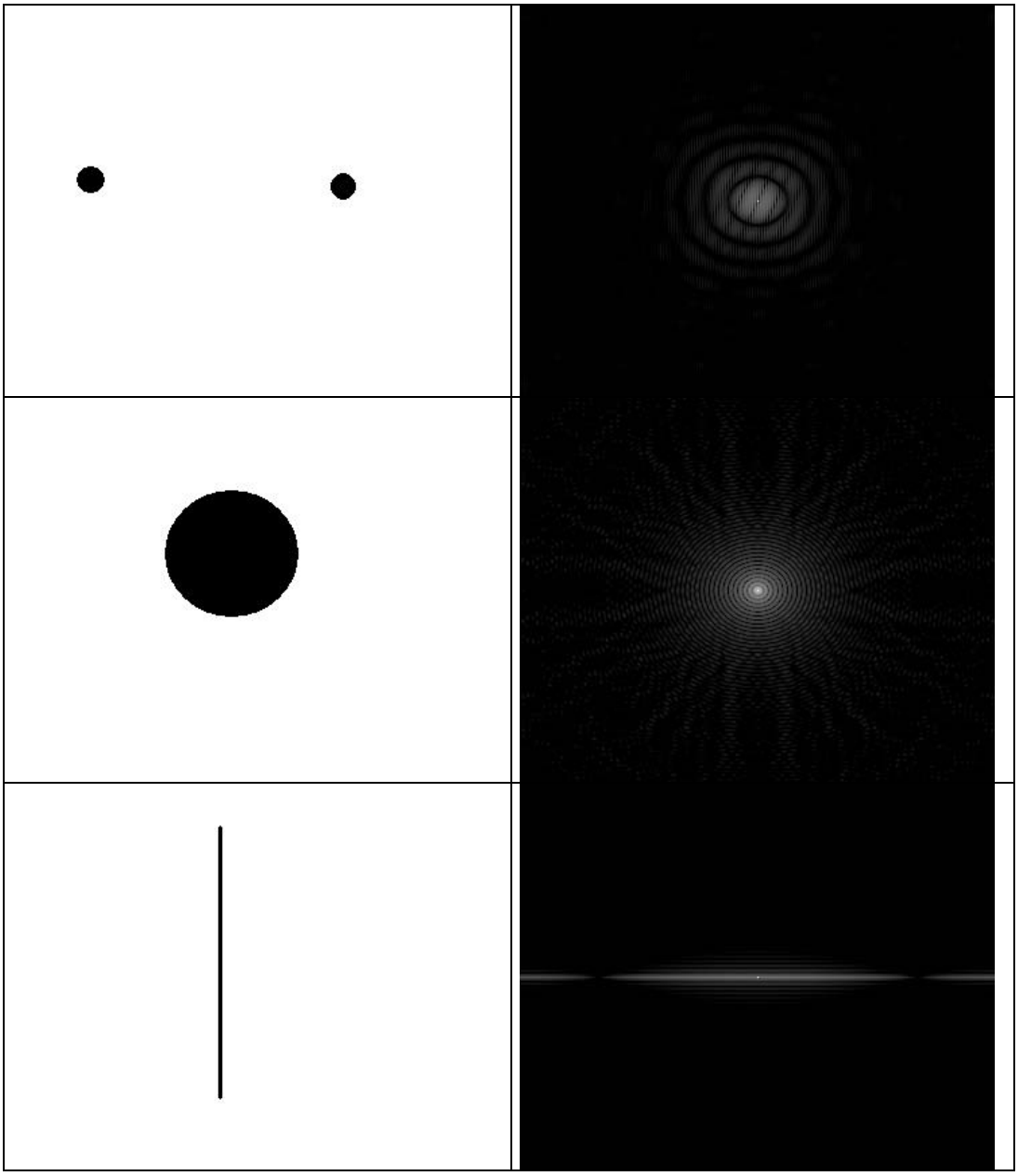
Finalment, s'ha implementat una eina per adjuntar els dos documents txt en un mateix document de text. El fitxer de text de sortida conté les característiques d'àudio i vídeo. L'eina utilitzada és la mostrada en la Fig. 1.1.10.

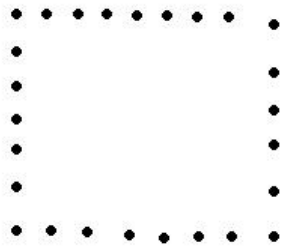
**Fig. 1.1.10** Eina per adjuntar característiques d'àudio i vídeo

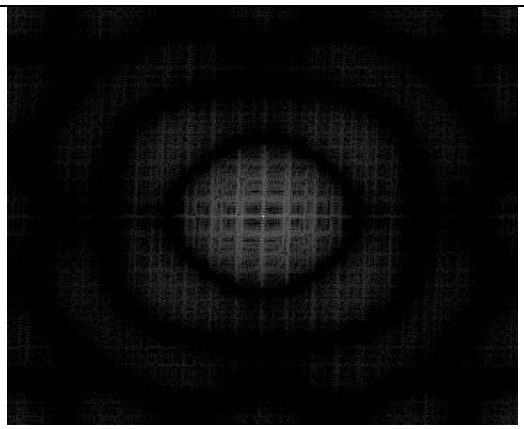
## 1.2. Proves de verificació del funcionament de la FFT

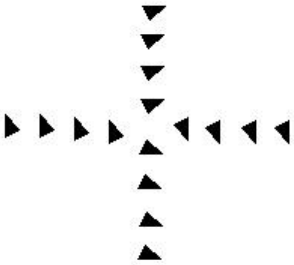
Per verificar si la FFT implementada en el programa funcionava correctament, s'han fet proves de verificació comparant les FFT obtingudes amb unes de mostra extretes d'un llibre de processat d'imatge (veure [8] ) .

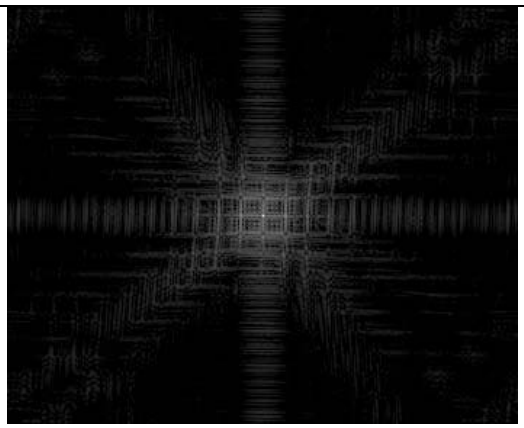
**Taula. 1.2.1** Imatges i les seves transformades de Fourier











## 1.3. Fitxer de característiques del descriptor de textura reduït

| TC | MeanHTD | DsvHTD | GabFil01 | GabFil02 | GabFil03 | GabFil04 | GabFil05 | DevFil01 | DevFil02 | DevFil03 | DevFil04 | DevFil05 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ----------FRAME LEVEL FEATURES---------- | | | | | | | | | | | | |
| 0 | 86.52468 | 36.97545 | 16.39147 | 19.26203 | 9.321081 | 4.225594 | 1.987876 | 1.351655 | 2.123284 | 1.730200 | 1.308332 | 0.973775 |
| 1920 | 86.52549 | 36.98706 | 16.63529 | 19.10345 | 9.202487 | 4.162577 | 1.958027 | 1.375857 | 2.132502 | 1.732507 | 1.308042 | 0.973532 |
| 3840 | 86.55519 | 37.02353 | 17.11966 | 18.83044 | 8.991141 | 4.058042 | 1.906967 | 1.429610 | 2.152653 | 1.738246 | 1.308574 | 0.973717 |
| 5760 | 87.03869 | 36.60760 | 16.75701 | 19.10066 | 9.184995 | 4.068614 | 1.903421 | 1.409303 | 2.148010 | 1.775300 | 1.317632 | 0.969184 |
| 7680 | 86.87052 | 36.60564 | 17.03107 | 18.90711 | 9.059088 | 4.022310 | 1.880191 | 1.442864 | 2.155012 | 1.778566 | 1.321294 | 0.970165 |
| 9600 | 86.95484 | 36.79758 | 16.08389 | 19.53381 | 9.467808 | 4.209103 | 1.972409 | 1.327378 | 2.134745 | 1.763519 | 1.315059 | 0.970108 |
| 11520 | 86.97174 | 36.81894 | 16.35263 | 19.34704 | 9.346362 | 4.153650 | 1.945840 | 1.355403 | 2.141194 | 1.765631 | 1.315900 | 0.971283 |
| 13440 | 87.06794 | 36.63548 | 16.28110 | 19.44922 | 9.321702 | 4.155257 | 1.950542 | 1.374588 | 2.154362 | 1.769224 | 1.314061 | 0.970408 |
| 15360 | 87.13139 | 36.66203 | 16.81386 | 19.11856 | 9.090323 | 4.054065 | 1.901826 | 1.426906 | 2.164342 | 1.769112 | 1.312894 | 0.970916 |
| 17280 | 101.9824 | 34.00955 | 7.886345 | 18.88268 | 15.97124 | 9.139493 | 4.750706 | 0.524561 | 1.006167 | 1.161302 | 1.025271 | 0.815774 |
| 19200 | 101.0224 | 34.96258 | 7.991867 | 18.73027 | 15.86061 | 9.314931 | 4.773523 | 0.527900 | 1.001424 | 1.156640 | 1.046464 | 0.815061 |
| 21120 | 87.66546 | 38.52551 | 12.01548 | 20.50701 | 12.13117 | 5.812313 | 2.797559 | 0.936882 | 1.780678 | 1.690017 | 1.332682 | 1.013750 |
| 23040 | 87.98925 | 38.64241 | 11.77929 | 20.70465 | 12.19044 | 5.844593 | 2.813347 | 0.934292 | 1.788239 | 1.693587 | 1.335942 | 1.014434 |
| 24960 | 89.87889 | 37.14002 | 12.59133 | 20.77656 | 11.56115 | 5.358640 | 2.461629 | 1.019494 | 1.899167 | 1.723149 | 1.338862 | 0.962054 |
| 26880 | 89.85611 | 37.17073 | 12.40393 | 20.87809 | 11.69440 | 5.407289 | 2.486764 | 1.010930 | 1.901541 | 1.727058 | 1.343303 | 0.967466 |
| 28800 | 86.54582 | 40.10445 | 12.67250 | 20.60533 | 11.64080 | 5.375264 | 2.517839 | 1.040698 | 1.858889 | 1.713046 | 1.306208 | 0.950978 |
| 30720 | 86.27074 | 40.16633 | 12.49331 | 20.69332 | 11.76499 | 5.392600 | 2.528277 | 1.039063 | 1.868825 | 1.729385 | 1.311680 | 0.956671 |
| 32640 | 77.43237 | 41.69498 | 14.79392 | 19.87195 | 10.31492 | 4.748681 | 2.138031 | 1.291468 | 2.157646 | 1.859580 | 1.437045 | 1.011048 |
| 34560 | 77.81721 | 41.90206 | 14.58628 | 19.96552 | 10.44130 | 4.786265 | 2.152789 | 1.287021 | 2.158324 | 1.870879 | 1.442345 | 1.013161 |
| 36480 | 76.12116 | 46.18795 | 16.98278 | 18.67857 | 9.321414 | 4.236371 | 1.800691 | 1.516852 | 2.298715 | 1.962666 | 1.500856 | 0.978079 |
| 38400 | 76.29089 | 46.40338 | 16.57189 | 18.92629 | 9.483458 | 4.334679 | 1.840493 | 1.486324 | 2.285506 | 1.956537 | 1.502375 | 0.978678 |
| 40320 | 81.39486 | 51.63818 | 14.84372 | 19.90226 | 10.22386 | 4.782030 | 2.044484 | 1.284777 | 2.155807 | 1.850961 | 1.463555 | 0.959084 |
| 42240 | 81.48002 | 51.78500 | 14.95262 | 19.84300 | 10.17716 | 4.764023 | 2.032967 | 1.297789 | 2.156765 | 1.855110 | 1.468140 | 0.958047 |
| 44160 | 80.38436 | 51.31632 | 15.09544 | 20.31250 | 9.761128 | 4.366007 | 1.974375 | 1.347272 | 2.325937 | 1.905210 | 1.430603 | 1.024571 |