TREBALL DE FI DE CARRERA

TÍTOL: Separació de *shots* de vídeo amb anàlisis multimodal

TITULACIÓ: Enginyeria Tècnica de Telecomunicació, especialitat Sistemes de Telecomunicació

AUTOR: Pere Palou Llobera

DIRECTOR: Francesc Tarrés Ruiz

DATA: 20 de gener de 2006

Títol: Separació de *shots* de vídeo amb anàlisis multimodal

Autor: Pere Palou Llobera

Director: Francesc Tarrés Ruiz

Data: 20 de gener de 2006

Resum

La indexació i la recuperació de vídeo en format digital és una de les àrees del tractament digital de senyals audiovisuals en les quals s'està desenvolupant una gran activitat. La quantitat d'informació audiovisual digital disponible en bases de dades està creixent de forma espectacular gràcies al desenvolupament tecnològic en la societat de la informació i la comunicació en els últims anys. Per aquesta raó, l'accés a les dades audiovisuals ha de ser el més senzill i ràpid possible per a estalviar temps i recursos.
Per això es necessiten **eines automàtiques de segmentació**, que separin una seqüència de vídeo en els seus *shots* elementals.

S'han implementat dos descriptors de color basats en histogrames definits en l'estàndard MPEG-7, el ***Scalable Color Descriptor*** (SCD), que extreu els bins de l'histograma de l'espai de color HSV, i el ***Group-of-Frames Descriptor*** (GoF), que s'utilitza per a representar el contingut de cada *shot* detectat mitjançant l'acumulació de tres histogrames diferents.
Una vegada extretes les característiques de color, es calculen mesures de distància L_2 entre *frames* consecutius que proporcionen la informació necessària per a, aplicant algorismes basats en llindars temporals adaptatius, detectar els *shots* (*hard cuts*) d'una seqüència de vídeo.

Es presenten un conjunt de resultats per a tots els gèneres de vídeo inclosos en la base de dades segmentada manualment. Aquests resultats s'avaluen a partir de la mesura de distància L_2 entre *frames* consecutius per als **paràmetres estadístics** μ i σ del canal HSV i, per altra banda, a partir de la mesura de distància L_2 entre *frames* consecutius per als **bins de l'histograma** extret pel SCD.

Recall i *Precision* mesuren la qualitat de les deteccions. Per a la valoració global del gènere de vídeo s'obtenen els següents resultats:

$$\text{Recall}_{bins}\ (97{,}29\%) > \text{Recall}_{\mu,\sigma}\ (92{,}69\%)$$
$$\text{Precision}_{bins}\ (78{,}92\%) < \text{Precision}_{\mu,\sigma}\ (86{,}51\%)$$

Title: Video shots separation using multimodal analysis

Author: Pere Palou Llobera

Director: Francesc Tarrés Ruiz

Date: January, 20th 2006

Overview

The indexing and retrieval of digital video is one of the most relevant areas in digital video processing.

The available amount of audiovisual information in data bases is growing in spectacular way due to technology development in the last years. For this reason, the access of audiovisual data as fast and simple as possible to save time and resources.

Automatic tools of segmentation, which segment a video sequence into their elementary shots, are needed.

Two colour-histogram based descriptors have been developed, Scalable Color Descriptor (SCD), that extracts HSV histogram bins, and Group-of-Frames Descriptor (GoF), which represents the content of each shot detected with the aggregation of three different histograms.

When colour features are extracted, distance measure L_2 between consecutive frames is computed. L_2 gives the information for detecting shots (hard cuts) with an adaptative thresholding algorithm.

A set of results for all types of video included in the database manually segmented. This set of results are evaluated in two ways; the first one is from distance measure L_2 for HSV channel μ and σ statistical parameters, and the other way is from distance measure L_2 for histogram bins extracted with SCD.

Recall and Precision measure the detection's quality. For a mixed set of video types this results are obtained:

$$\text{Recall}_{bins}\ (97{,}29\%) > \text{Recall}_{\mu,\sigma}\ (92{,}69\%)$$
$$\text{Precision}_{bins}\ (78{,}92\%) < \text{Precision}_{\mu,\sigma}\ (86{,}51\%)$$

A n'Antònia, en Miquel i n'Aina, per la seva incondicional estima.

Als germans dels dijous nit, als companys de l'EPSC
i als amics de convivència.

A en Toni i en Francesc, per l'ajuda, les mostres d'humor
i les petites empentes que m'han donat.

I a qui m'omple cada dia estant al meu costat.

"Ocupa't i despreocupa't"

ÍNDEX

INTRODUCCIÓ

El desenvolupament tecnològic en la societat de la informació i la comunicació en els últims anys, com és el creixement d'Internet, ha provocat l'augment considerable de la quantitat de bases de dades audiovisuals en format digital, que encara ara està creixent.

Per a veure'n un exemple, la **Fig. 0.1** mostra el consum anual de bytes als EUA l'any 2000.

Total for 70M households	**~230 Exabyte/year**
Television	**94%**
Radio	1.7%
Recorded Music	0.4%
Newspaper	0.0003%
Books	0.0002%
Magazines	0.0002%
Home video	**3.3%**
Video games	**0.6%**
Internet	0.0003%
	[Source: UC Berkeley: How much Information]

Fig. 0.1 Consum anual de bytes als EUA (any 2000)

En setanta milions de llars el consum va ser aproximadament de 230 Exabytes (10^{18} bytes). Per a tenir una referència de la quantitat d'informació que això comporta es pot comparar amb els 2 Exabytes de la informació generada l'any 1999 o amb els 5 Exabytes de totes les paraules parlades pels éssers humans.

Amb aquestes dades és senzill veure la dificultat de manegar el contingut audiovisual. Per això es necessiten eines automàtiques de segmentació.

A grans trets, l'anàlisi de contingut de vídeo implica una sèrie d'etapes [1]. La primera etapa és la **segmentació**. Per a un vídeo, això significa que s'ha de separar la seqüència sencera fins a obtenir-ne els seus *shots* elementals.

Un ***shot*** és una seqüència de *frames* (imatges que componen un vídeo) capturada per una càmera en una acció contínua en el temps i l'espai. Un grup de *frames* (GoF) que té característiques visuals consistents (com és el color, la textura i el moviment) equival a un *shot*.
Típicament, la direcció d'una càmera i el seu angle de visió defineix un *shot*; quan una càmera enfoca la mateixa escena des d'angles diferents o a diferents regions d'una escena des del mateix angle de visió, s'observen *shots* diferents.

L'etapa posterior a la segmentació és la classificació de *shots* en categories predefinides de diferent nivell semàntic (baix / mig / alt). Un exemple de la jerarquia dels nivells semàntics seria un vídeo corresponent a esport (baix nivell), de futbol (nivell mig) i un llançament de penal (alt nivell).
Altres etapes de l'anàlisi són la indexació de contingut en bases de dades i el resum del contingut audiovisual per a la recuperació de seqüències de vídeo.

L'objectiu d'aquest Treball de Fi de Carrera és l'anàlisi del contingut audiovisual (AV) digital per a detectar i separar automàticament els *shots* de segments de vídeo.
Mitjançant eines descriptives que siguin capaces d'extreure característiques d'àudio i de vídeo, es podran desenvolupar mètodes per a la detecció i separació automàtica de *shots* de seqüències de vídeo.

Les tècniques de segmentació de contingut AV mitjançant anàlisi **multimodal** engloba la conjunció de, per una part, l'anàlisi de contingut d'àudio i, per altra part, l'anàlisi de contingut de vídeo. Seguidament s'ha de desenvolupar una etapa d'integració dels dos anàlisis.

Cal dir, que aquest treball s'ha enfocat exclusivament en l'etapa de descripció de continguts visuals. S'implementaran dos descriptors de color, el *Scalable Color Descriptor* (SCD) i el *Group of Frames/Pictures Descriptor* (GoF), definits en l'estàndard de descripció de contingut multimèdia, MPEG-7.
Aquests descriptors estan basats en histogrames de color, els quals capturen la distribució global del color.

Posteriorment, es calcularan diverses mesures de similitud entre *frames* consecutius. Aquestes mesures donen la informació necessària per a, aplicant algorismes basats en llindars temporals adaptatius, detectar els *shots* d'una seqüència de vídeo.

Per a la comprovació del funcionament dels descriptors i dels algorismes es realitza una costosa tasca manual de segmentació i descripció de material AV. Aquesta base de dades és útil per a la comprovació del funcionament de l'algorisme de detecció de *shots*.

Aquesta memòria està organitzada en quatre capítols i un annex, els quals es detallen a continuació:

Al primer capítol es presenten detalladament els conceptes teòrics dels mètodes utilitzats per a la de **detecció de *shots* de vídeo**.

Al segon capítol es duu a terme una explicació de la **segmentació i la descripció de material audiovisual** seleccionat, així com el procediment d'etiquetatge que s'ha seguit.

Al tercer capítol s'expliquen els **resultats** obtinguts a partir dels algorismes que s'han implementat i poder així fer una avaluació de la seva eficiència.

El quart capítol té per objectiu l'explicació de les **conclusions** obtingudes a partir del desenvolupament dels algorismes i de l'aplicació que els sustenta i a partir dels resultats extrets. També s'inclou un apartat sobre possibles **línies futures** d'investigació.

Finalment, l'annex mostra l'**aplicació visual** desenvolupada per a aplicar els algorismes de detecció de *shots* detallats en el primer capítol, així com les eines utilitzades per a implementar el sistema.

CAPÍTOL 1. DETECCIÓ DE *SHOTS* DE VÍDEO

1.1. Introducció

La indexació i la recuperació de vídeo en format digital és una de les àrees del tractament digital de senyals audiovisuals en les quals s'està desenvolupant una gran activitat. La quantitat d'informació audiovisual digital disponible en bases de dades està creixent de forma desmesurada (veure **Fig. 0.1**).

Per aquesta raó, l'accés a les dades audiovisuals ha de ser el màxim de senzill i ràpid possible per a estalviar temps i recursos. Per a dur a terme aquest objectiu, és necessària una etapa de segmentació temporal de vídeo mitjançant la detecció dels límits entre *shots* que el componen.

Com s'ha explicat a la part introductòria d'aquest document, un ***shot*** és una seqüència de *frames* capturada per una càmera en una acció contínua en el temps i l'espai.

La **Fig. 1.1** il·lustra una sèrie de canvis *shot* dins una seqüència de vídeo.

Fig. 1.1 Exemple d'una seqüència de vídeo d'un anunci publicitari amb 8 *shots*

Els tipus de *shots* es poden classificar de la següent forma:

- *Hard Cuts*
  Són canvis de càmera bruscs (*camera breaks*) que succeeixen en la transició entre dos *frames*.

- *Soft Cuts*
  Són transicions graduals entre diferents *shots*, la durada d'aquestes transicions pot anar aproximadament des dels 5 *frames* als 30 *frames*. Hi ha diferents tipus de transicions graduals:

- *Wipe:* Transició gradual espacial on una línia, una circumferència en expansió, un objecte...es mou a través de la pantalla
- *Fade-in:* Comença amb un *frame* negre i apareix gradualment la imatge del següent *shot*.
- *Fade-out:* És l'oposat d'un *fade-in*.
- *Dissolve:* És la superposició d'un *fade-out* sobre un *fade-in*.

A causa dels avanços de la tecnologia en general, i dels equips audiovisuals en particular, cada cop hi ha més transicions graduals amb efectes especials complicats que dificulten molt la seva detecció automàtica.

La **Fig. 1.2** mostra un exemple de transició gradual amb efectes especials entre *shots* d'una seqüència de vídeo.

Fig. 1.2 Exemple de transició gradual amb efectes especials

Els *shots* són les unitats elementals d'una seqüència de vídeo. Conèixer la posició temporal exacta de les transicions entre *shots* és realment útil i té un ampli ventall d'**aplicacions** en diferents àrees; principalment en la recuperació i indexació de vídeo:

- Recuperació de vídeo

  A causa de l'avanç tecnològic global, concretament en Internet, en multimèdia i en l'estandardització de tècniques de compressió de vídeo, una quantitat desmesurada de material de vídeo és digitalitzat. En aplicacions en les quals es volen trobar vídeos a partir d'una demanada (recerca de material en una videoteca d'una cadena de TV) mitjançant la comparació de similituds entre aquesta demanda i bases de dades, es fa necessari representar temporalment seqüències de vídeo.

- Indexació de vídeo

  El principal avantatge de detectar *shots* automàticament per a la indexació de vídeo és que es pot evitar la segmentació i descripció manual d'un conjunt de material audiovisual, que és un treball que requereix molt de temps de dedicació (el qual s'ha pogut experimentar en la realització d'aquest treball).
  Amb l'ajuda de descriptors de contingut visual, com és el cas de les eines incloses en l'estàndard MPEG-7, és possible indexar els segments de vídeo automàticament i descriure'n el seu contingut.

Com s'ha comentat, per a detectar automàticament els *shots* de segments de vídeo, es necessiten eines que descriguin el seu contingut de forma automàtica.
Per a fer-ho s'han implementat dos descriptors de color definits en l'estàndard MPEG-7, el *Scalable Color Descriptor* i el *Group-of-Frames Descriptor*, els dos basats en histogrames de color, que capturen la distribució global del color.

Tot seguit, s'han calculat diverses mesures de similitud entre *frames* consecutius per a detectar els canvis de càmera mitjançant algorismes basats en llindars adaptatius.

1.2. Estàndard MPEG-7

L'estàndard MPEG-7, ISO/IEC 15938, formalment anomenat *Multimedia Content Description Interface*, defineix un conjunt d'eines de descripció de contingut audiovisual. L'objectiu de l'estàndard és facilitar l'accés al contingut audiovisual, el qual implica emmagatzemament, identificació, filtratge, recerca i recuperació eficient de multimèdia.

Abans de l'MPEG-7, les dades audiovisuals eren vistes majoritàriament com sèries de bits. Nomes la descodificació d'aquests bits podia donar alguna informació sobre el contingut de les dades i el tractament que es podia aplicar sobre aquestes.
El procés de descodificació implica, generalment, operacions complexes i de gran reserva de memòria i requereix grans amplades de banda en entorns de desenvolupament en línia, la qual cosa no és factible.

Amb l'ús dels seus descriptors, l'estàndard MPEG-7 proporciona la possibilitat d'aconseguir la informació sobre les dades audiovisuals sense la necessitat d'interpretar la descodificació de les dades.

L'estàndard s'estructura en diferents parts (veure [2]). En la *Part 3 – Visual* és on s'especifiquen els descriptors visuals. En el següent apartat es detallen aquests descriptors, i en concret els utilitzats per a desenvolupar els algorismes per a la detecció de *shots* i per a la descripció del seu contingut.

1.3. Descriptors Visuals

El principal objectiu de la part visual de l'estàndard MPEG-7 és proporcionar descripcions estandarditzades d'imatges emmagatzemades o en *streaming* que ajudin a usuaris o a aplicacions a identificar, categoritzar o filtrar imatges o vídeos.

Els descriptors visuals de l'estàndard MPEG-7 descriuen el contingut audiovisual. Per a imatges i vídeo, es descriu el seu contingut semàntic amb

diferents característiques com el color, la textura, la forma i el moviment. L'estàndard inclou descriptors específics per a cadascuna d'aquestes característiques. Aquests, s'exposen a continuació [2]:

1.3.1. Descriptors de Color

El color és una de les característiques visuals més àmpliament utilitzades en la recuperació d'imatge i vídeo. Les característiques de color són relativament invariants a l'angle de visió i a la translació i rotació de la imatge.
A l'estàndard hi ha sis descriptors de colors que representen diferents aspectes del color, els quals es presenten a continuació:

- *Color Space Descriptor (Color Quantization Descriptor)*
- *Dominant Color Descriptor*
- *Scalable Color Descriptor*
- *Group-of-Frame/Group-of-Picture Descriptor*
- *Color Structure Descriptor*
- *Color Layout Descriptor*

L'apartat 1.4. explica els diferents descriptors de color i els apartats 1.5 i 1.6 detallen els dos descriptors implementats en aquest treball, el *Scalable Color Descriptor* i el *Group-of-Frame/Group-of-Picture Descriptor*.

1.3.2. Descriptors de Textura

La textura es refereix als models (patrons) visuals que tenen o no propietats d'homogeneïtat, que resulten de la presència de múltiples colors o intensitats dins la imatge.
És una característica important d'una superfície visible virtual on hi ha la repetició, parcial o total, d'un model fonamental.
La descripció de textures en imatges amb els descriptors de l'estàndard, esmentats a continuació, proporcionen mitjanes molt potents per a la comparació de similituds i per a la recuperació.
A continuació s'esmenten els tres descriptors de textura inclosos en l'estàndard MPEG-7:

- *Homogeneous Texture Descriptor*
- *Texture Browsing Descriptor*
- *Edge Histogram Descriptor*

1.3.3. Descriptors de Forma

En moltes aplicacions de bases de dades d'imatges, la forma dels objectes d'una imatge proporcionen un potent tret visual per a la comparació de similituds. L'estàndard proporciona descriptors per a detectar regions o contorns en imatges, els quals són útils per a aplicacions com, per exemple, detectar els contorns de caràcters escrits en imatges binàries.

A continuació s'esmenten els tres descriptors de forma inclosos en l'estàndard MPEG-7:

- *Region-Based Shape Descriptor*
- *Contour-Shape Descriptor*
- *3-D Shape Descriptor*

1.3.4. Descriptors de Moviment

Tots els descriptors de color, textura i forma d'objectes poden ser utilitzats per a indexar imatges en seqüències de vídeo. La descripció de característiques de moviment en seqüències de vídeo pot proporcionar trets encara més potents respecte el seu contingut. Els descriptors de moviment desenvolupats a l'estàndard MPEG-7, capturen característiques de moviment (com són el moviment de càmera i el moviment d'objectes) en descripcions concises i efectives.
A continuació s'esmenten els quatre descriptors de moviment inclosos en l'estàndard MPEG-7:

- *Motion Activity Descriptor*
- *Camera Motion Descriptor*
- *Motion Trajectory Descriptor*
- *Parametric Motion Descriptor*

1.4. Descriptors de Color

El color és un atribut important d'informació visual. La informació de color facilita la nostra vida diària, per exemple, quan es veu un senyal lluminós en un semàfor o en la identificació d'un equip en un esdeveniment esportiu.

Una gran quantitat de recerca s'ha desenvolupat en diferents aspectes de la caracterització del color en base a la seva percepció, coherència i distribució espacial. L'estàndard MPEG-7 reflexa el resultat d'aquesta investigació en forma de descriptors.
Diferents factors influencien la selecció d'aquests descriptors de color, que inclou:

- Qualitat
  L'habilitat de caracteritzar la similitud perceptiva del color, la qual ha estat avaluada a partir del funcionament dels descriptors en la comparació entre segments de vídeos basats en característiques de color.

- Simplicitat
  Baixa complexitat de les tècniques d'extracció i comparació dels descriptors, ja que els sistemes que els utilitzin han de ser capaços de realitzar tasques de recerca i de recuperació a través de grans bases de dades de multimèdia, així com en dispositius portàtils amb potència computacional limitada.

- Codificació eficient
  La mida de les descripcions codificades, la qual cosa és important en la indexació i en la transmissió dels descriptors a través de xarxes d'amplades de bandes limitades.

- Escalabilitat
  L'escalabilitat, que indica la capacitat d'un sistema d'incrementar el rendiment de processament on s'hi afegeixen recursos, i la interoperabilitat dels descriptors.

Com s'ha esmentat al subapartat 1.3.1., existeixen sis descriptors [2] [4]. Tot seguit s'expliquen les bases de cadascun d'ells i posteriorment es dediquen dos apartats a detallar els dos descriptors implementats.

La **Fig.1.3** mostra un esquema que inclou tots els descriptors de color definits en l'estàndard MPEG-7.

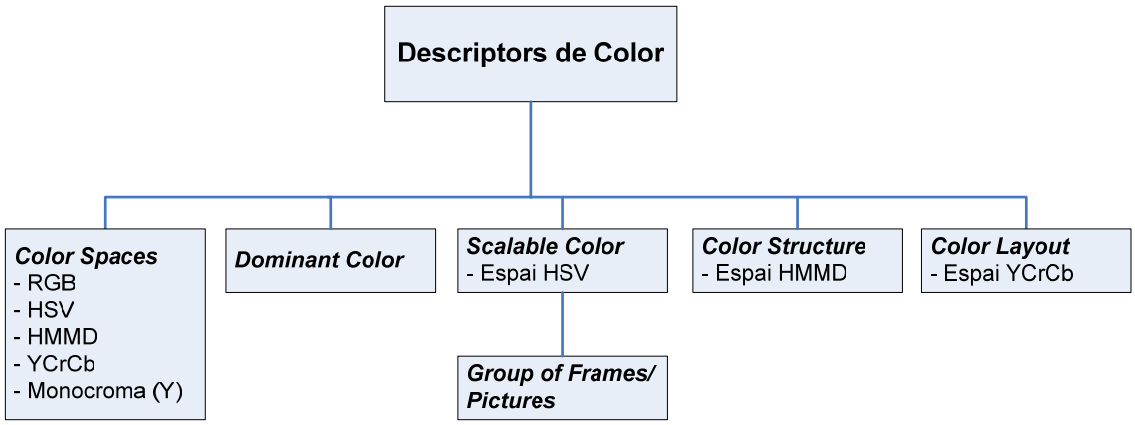


Fig.1.3 Descriptors de color de l'estàndard MPEG-7

- *Color Space Descriptor*

  Permet la selecció d'un espai de color per a utilitzar-lo en la descripció de contingut visual. El *Color Quantization Descriptor* especifica la partició de l'espai de color donat en *bins* discrets.
  Aquests dos descriptors s'utilitzen conjuntament amb altres descriptors de color, específicament, en el *Dominant Color Descriptor*. Els espais de color especificats són RGB, HSV, HMMD, YCbCr i monocroma (Y).

- *Dominant Color Descriptor*

  Permet l'especificació d'un curt nombre de valors de color, així com les seves propietats estadístiques, com la distribució i la variància. El seu objectiu és proporcionar una representació efectiva, compacta i intuïtiva dels colors presents en una regió o en una imatge.

- *Scalable Color Descriptor*

  Està basat en un histograma de color definit en l'espai de color *HSV (Hue-Saturation-Value)* amb una quantificació fixada de l'espai de color. Utilitza una codificació basada en els coeficients extrets d'una Transformada *Wavelet Haar*, la qual permet una representació escalable de descripció.

- *Group-of-Frame or Group-of-Picture Descriptor*

  És una extensió del *Scalable Color Descriptor* a un grup de *frames* en un vídeo o en una col·lecció d'imatges. Aquest descriptor es basa en l'agregació les propietats de color de les imatges individuals o dels *frames* d'un vídeo.

- *Color Structure Descriptor*

  Està basat en histogrames de color. Té com a objectiu la identificació de distribucions localitzades de color utilitzant petites finestres. L'estàndard defineix que per a garantir la interoperabilitat entre els descriptors, el *Color Structure Descriptor* està limitat a l'espai de color HMMD.

- *Color Layout Descriptor*

  Captura la disposició espacial de colors representatius a una graella superimposada a una regió o imatge. La representació està basada en els coeficients de la Transformada Cosinus Discreta (DCT).
  És un descriptor molt compacte i presenta una gran eficiència en aplicacions de ràpida recerca. Ës aplicable tant a imatges fixes com a segments de vídeo.

1.5. *Scalable Color Descriptor*

El *Scalable Color Descriptor* (SCD) és un histograma de color extret de l'espai de color HSV, i codificat mitjançant la Transformada *Wavelet Haar*.
Tot seguit, s'expliquen els conceptes teòrics per a entendre el funcionament del descriptor [2] [3].

1.5.1. Espais de color RGB i HSV

Un espai de color es defineix com un model de representació de color en termes de valors d'intensitat. Els espais més usats en sistemes de recuperació d'imatges són l'RGB i l'HSV.

Espai de color RGB

Està composat dels colors primaris *Red* (vermell), *Green* (verd) i *Blue* (blau). El model RGB utilitza el sistema de coordenades cartesià com es mostra en la **Fig. 1.4 (a)** (la diagonal des del negre (0,0,0) al blanc (1,1,1) representa l'escala de grisos). La **Fig. 1.4 (b)** mostra l'espai de color RGB.

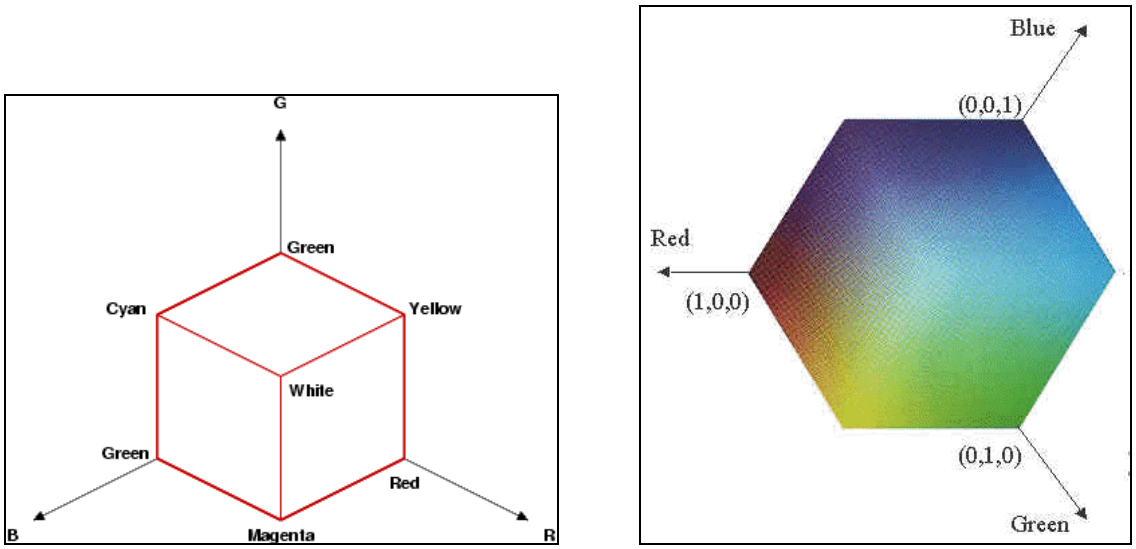


Fig. 1.4 (a) Sistema de coordenades RGB. (b) Espai de color RGB

Espai de color HSV

Consisteix en *Hue*, *Saturation & Value* (també anomenat HSB: *Hue-Saturation-Brightness)*. Aquest espai de color és un dels més usats a causa de les seves similituds amb la forma amb la qual l'ull humà tendeix a percebre el color.
L'espai es pot definir d'igual forma com un objecte cònic o com un objecte cilíndric (veure **Fig. 1.5**).

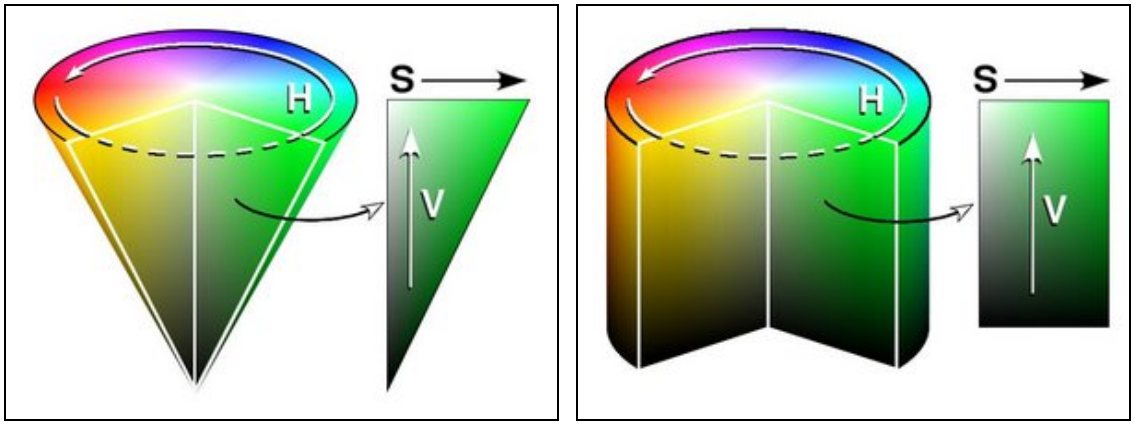


Fig. 1.5 Espais de colors HSV cònic i cilíndric

Cada component de l'espai de color especifica:

- *Hue*:
  - El **color**.
  - Es representa mitjançant angles de 0 a 360º, com es pot veure a la **Fig. 1.6**.

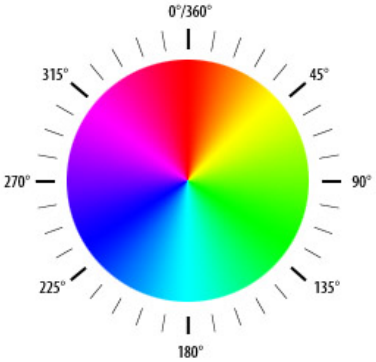


Fig. 1.6 Rang d'angles en el component *Hue* de l'espai de color HSV

- *Saturation*:
  - El percentatge de blanc (la **solidesa** del color).
  - El rang del canal s'especifica de 0 a 100%.
- *Value*:
  - La **intensitat** o lluminositat.
  - De la mateix forma que a *Saturation*, el rang del canal s'especifica de 0 a 100%.

La **Fig. 1.7** mostra dos exemples per il·lustrar els components de l'espai HSV. En el primer, el color de la regió de la imatges correspon a 209º (blau) del component H, mentre que al segon l'angle és de 61º (groc). Aquests angles fixen la variació dels components S i V.

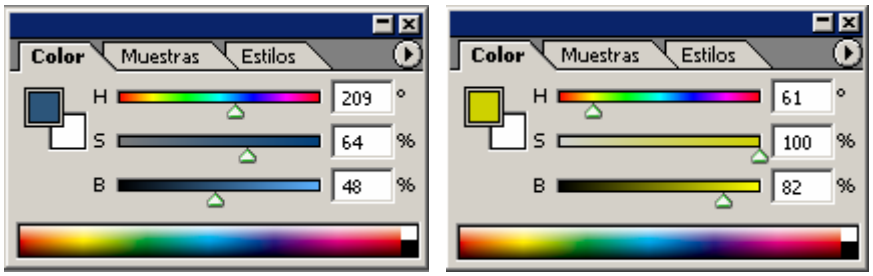


Fig. 1.7 Exemples de variació dels components de l'espai de color HSV

La **Fig. 1.8** mostra el codi per a fer la conversió de l'espai de color RGB a l'espai de color HSV:

Conversió de l'espai de color RGB a HSV

```
Max = max(R,G,B);
Min = min(R,G,B);
Value = Max;

if( Max == 0 ) then
    Saturation = 0; else
    Saturation = (Max - Min)/Max;
if( Max == Min ){
    Saturation = 0;     //Color acromàtic (blanc, negre o gris)
    Hue = 0; }          //Hue=0º (color vermell)
otherwise:
if( Max == R && G >= B )       Hue = 60*(G-B)/(Max-Min);
else if(Max == R && G < B )    Hue = 360 + 60*(G-B)/(Max-Min);
else if(G == Max)              Hue = 60*(2.0 + (B-R)/(Max-Min));
else                           Hue = 60*(4.0 + (R-G)/(Max-Min));
```

Fig. 1.8 Conversió de l'espai de color RGB a HSV

1.5.2. Histograma HSV

Un histograma és una representació bidimensional de la freqüència relativa amb la que apareix cada nivell de color o de gris en una imatge. Així, per a cada possible nivell de color o de gris en l'eix d'abcises, trobem el número de píxels de la imatge que prenen aquest nivell de color o de gris (veure **Fig. 1.9**). Per tant, l'histograma és una representació aproximada de la probabilitat de que un píxel prengui un determinat nivell de color en una imatge.

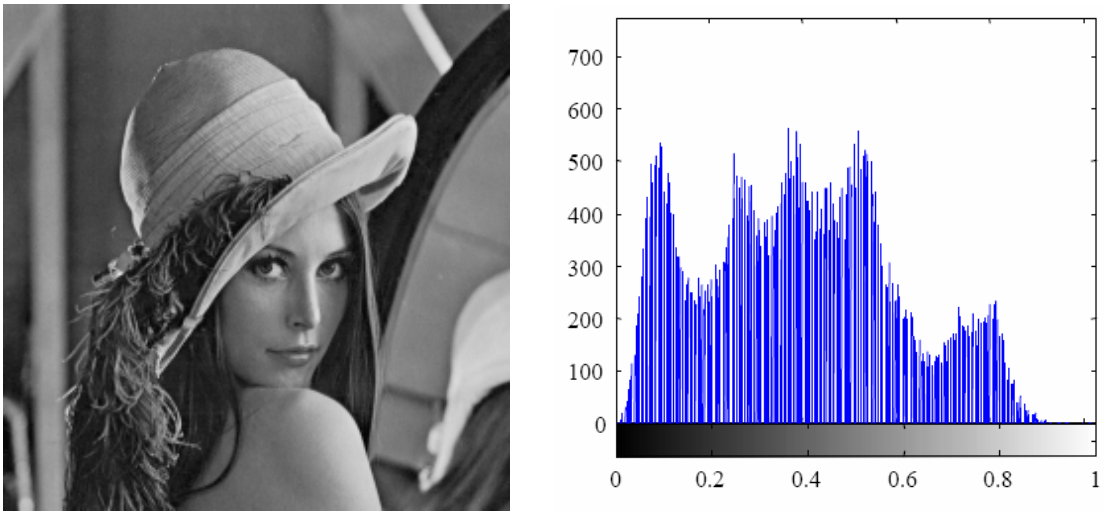


Fig. 1.9 Imatge en escala de grisos i el seu corresponent histograma normalitzat [0,1]

Extracció dels bins

Per a la implementació del SCD és necessari calcular l'histograma HSV de cada *frame* del segment de vídeo que s'està processant.
L'histograma de color que es crea és una reproducció tridimensional de l'espai de color HSV, és a dir, cadascun dels bins conté el número de píxels que tenen un nivell de color determinat corresponent a un nivell del component H, un nivell corresponent al component S i un corresponent al nivell V.
L'histograma es divideix en un determinat nombre porcions, definides per una quantificació, on cadascuna d'aquestes porcions correspon a un bin, el qual està associat a un nivell de color (H, S, V).
Cada component de l'histograma de color s'ha quantificat amb un número determinat de bits. Les quantificacions utilitzades per al descriptor són:

- 16 bins → 4:2:2
    - Canal H: 2 bits → $2^{2\ bits}$ = 4 bins
    - Canal S: 1 bit → $2^{1\ bit}$ = 2 bins
    - Canal V: 1 bit → $2^{1\ bit}$ = 2 bins

  En total s'utilitzen 4 bits per a aquesta quantificació, la qual cosa suposa que es podrà representar l'histograma amb $2^{4\ bits}$ = 16 bins, és a dir, 16 possibles nivells de color diferents.

- 32 bins → 8:2:2
    - Canal H: 3 bits → $2^{3\ bits}$ = 8 bins
    - Canal S: 1 bit → $2^{1\ bit}$ = 2 bins
    - Canal V: 1 bit → $2^{1\ bit}$ = 2 bins

  En total s'utilitzen 5 bits per a aquesta quantificació, la qual cosa suposa que es podrà representar l'histograma amb $2^{5\ bits}$ = 32 bins.

- 64 bins → 8:2:4
    - Canal H: 3 bits → $2^{3\ bits}$ = 8 bins
    - Canal S: 1 bit → $2^{1\ bit}$ = 2 bins
    - Canal V: 2 bits → $2^{2\ bits}$ = 4 bins

  En total s'utilitzen 6 bits per a aquesta quantificació, la qual cosa suposa que es podrà representar l'histograma amb $2^{6\ bits}$ = 64 bins.

- 128 bins → 8:4:4
    - Canal H: 3 bits → $2^{3\ bits}$ = 8 bins
    - Canal S: 2 bits → $2^{2\ bits}$ = 4 bins
    - Canal V: 2 bits → $2^{2\ bits}$ = 4 bins

  En total s'utilitzen 7 bits per a aquesta quantificació, la qual cosa suposa que es podrà representar l'histograma amb $2^{7\ bits}$ = 128 bins.

- 256 bins → 16:4:4

    - Canal H: 4 bits → $2^{4\ bits}$ = 16 bins
    - Canal S: 2 bits → $2^{2\ bits}$ = 4 bins
    - Canal V: 2 bits → $2^{2\ bits}$ = 4 bins

    En total s'utilitzen 8 bits per a aquesta quantificació, la qual cosa suposa que es podrà representar l'histograma amb $2^{8\ bits}$ = 256 bins.

Com s'ha dit al principi de l'apartat 1.5., l'extracció del descriptor consta de:

- Càlcul dels bins de l'histograma en l'espai de color HSV. S'escull la quantificació de 256 bins, amb 16 bins per al component H, 4 bins per al component S i 4 bins per al component V.
- Transformació dels bins extrets mitjançant la Transformada *Wavelet Haar* (veure 1.5.3.).

Extracció de paràmetres estadístics del canal HSV

A banda de l'extracció dels bins de l'histograma, es calculen, per a cada canal de l'espai de color HSV, la seva mitjana aritmètica (μ) i la desviació estàndard (σ). Això servirà per a tenir altres característiques per a la detecció de *shots*.

La **mitjana aritmètica (μ)** és un paràmetre estadístic de primer ordre, per a un conjunt de N mostres es defineix com:

$$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \qquad \textbf{(1.1)}$$

On,

N: nombre de mostres
x: mostra

La **variància (σ^2)** és un paràmetre estadístic de segon ordre que mesura la dispersió de la mostra x al voltant de la seva mitjana μ, és a dir, mesura si un conjunt de N mostres són, en general, properes o llunyanes a la seva mitjana. La **desviació estàndard (σ)** és l'arrel quadrada de la variància i es defineix de la següent manera:

$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2} \qquad \textbf{(1.2)}$$

On,

N: nombre de mostres
x: mostra
μ: mitjana del conjunt de N mostres

També es calcula la μ i la σ del canal HSV, a partir de la combinació dels tres canals H,S i V, de la següent forma:

$$\mu_{HSV} = \frac{\mu_H + \mu_S + \mu_V}{3} \qquad \sigma_{HSV} = \sqrt[3]{\sigma_H \cdot \sigma_S \cdot \sigma_V} \qquad \textbf{(1.3)}$$

1.5.3. Transformada *Wavelet Haar* (HWT)

Imatge 2D

La idea bàsica de la HWT aplicada sobre una imatge (2D) és la descomposició piramidal en sub-bandes aplicant-hi vàries etapes de delmació concentrant la seva informació en una zona determinada.
Esquemàticament es pot apreciar millor la seva funcionalitat mitjançant la **Fig. 1.10** que mostra el diagrama de blocs de la primera etapa de la transformació.

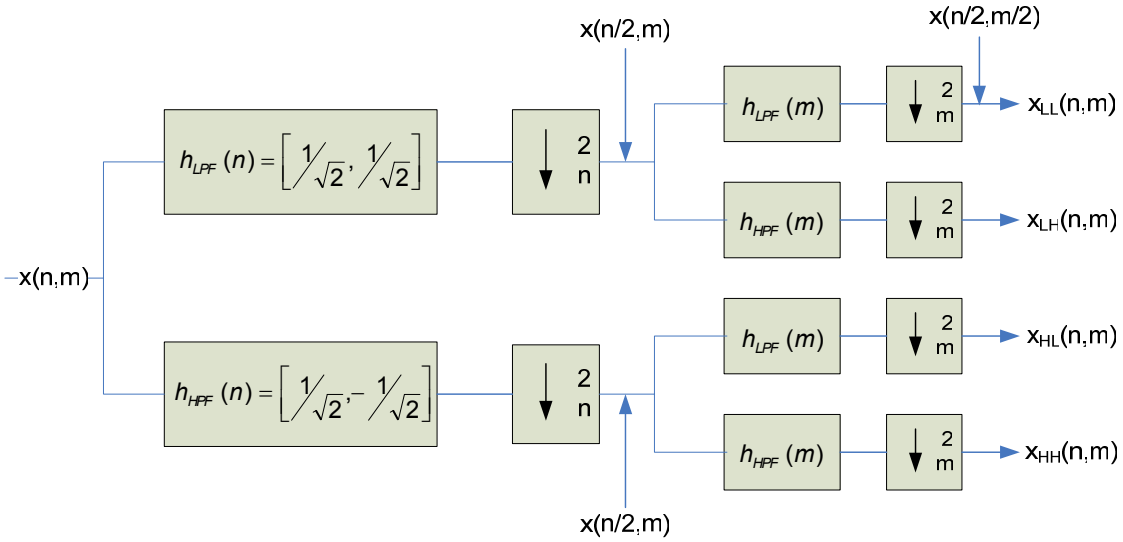


Fig. 1.10 Diagrama de blocs de la primera etapa de la transformació

En primer lloc, la imatge x(n,m) és descomposta en dues branques.
En la branca superior, s'hi aplica un filtre passa baixes freqüències, $h_{LPF}(n)$, en la direcció n (files → direcció horitzontal) de la imatge; tot seguit es delma per 2 la imatge en la mateixa direcció en què s'ha aplicat el filtratge.
La branca inferior és exactament igual que la superior amb la diferència del filtre, $h_{HPF}(n)$, que deixa passar les altes freqüències. A aquest punt tenim la imatge x(n,m) amb la meitat de files.

L'efecte que provoca el filtre LPF en direcció horitzontal és un suavitzat de la imatge en aquesta direcció. Per altra banda, l'efecte del filtre HPF en direcció horitzontal és l'emfatització dels contorns verticals de la imatge, ja que s'elimina

la informació dels contorns (canvis bruscs → alta freqüència) en aquesta direcció, quedant només la informació en direcció vertical.
Se segueix el mateix procediment per a la direcció m (columnes → direcció vertical). Els efectes dels filtres LPF i HPF en aquesta direcció seran els oposats als descrits per a la direcció horitzontal (files).

Al final de l'etapa, cada sub-imatge té la meitat de files i de columnes (x(n/2,m/2)), és a dir, de quatre vegades més petita que la original.

El resultat de la transformada és una imatge descomposada en 4 sub-imatges de mida 4 vegades inferior cadascuna (veure **Fig. 1.11**):

- La imatge resultant $x_{LL}(n,m)$ correspon a la imatge superior esquerra. Conté la informació de la imatge x(n,m) original.
- La imatge $x_{LH}(n,m)$ correspon a la imatge superior dreta. Té emfatitzats els contorns horitzontals.
- La imatge $x_{HL}(n,m)$ correspon a la imatge inferior esquerra. Té emfatitzats els contorns verticals.
- La imatge $x_{HH}(n,m)$ correspon a la imatge inferior dreta. Té emfatitzats els contorns diagonals.

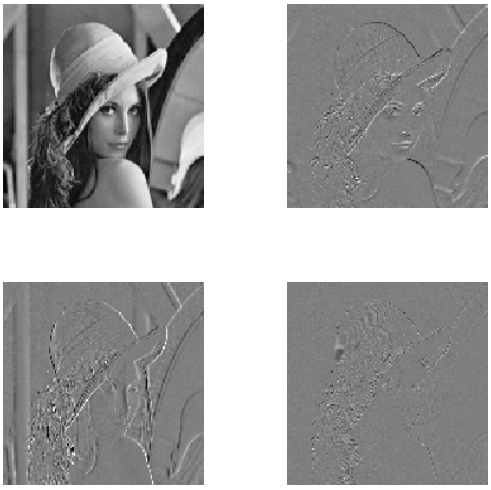


Fig. 1.11 Imatge resultant de l'aplicació d'una etapa de la HWT

La imatge resultant, formada pel conjunt de les quatre sub-imatges, es codifica entròpicament, aconseguint comprimir la seva informació.
A partir d'aquí, es realitzarien tantes etapes com es volgués comprimir la imatge, per exemple, per a la posterior transmissió. L'entrada de la següent etapa és la sortida de l'etapa anterior, és a dir, la imatge $x_{LL}(n,m)$, que és la que conté la informació original.
En el cas de voler reconstruir la imatge que s'ha codificat mitjançant aquesta transformada, es realitzaria el procés en ordre invers.

Histograma HSV

Un cop explicat el concepte bàsic de la HWT aplicat a una imatge 2D, s'explica la transformada en el domini dels bins de l'histograma HSV [2] [3].

L'entrada a la HWT són els 256 bins corresponents a les porcions del cilindre tridimensional que constitueix l'histograma de l'espai de color HSV (veure **Fig. 1.12**). A partir d'aquí, s'apliquen una sèrie d'etapes amb un ordre determinat: filtratge i delmació dels bins en direcció H, en direcció D, en direcció V i en direcció H un altre cop.

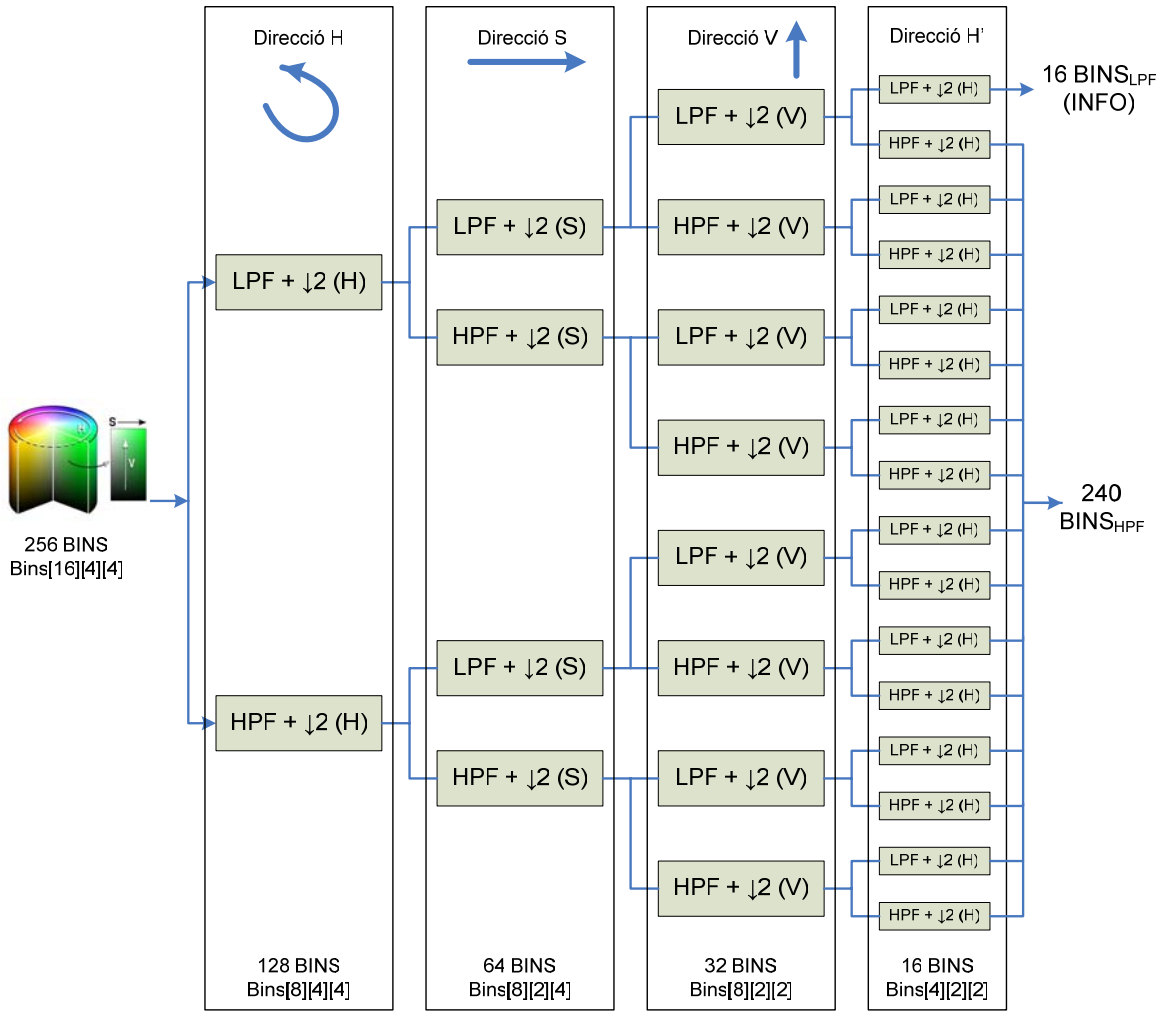


Fig. 1.12 Diagrama de blocs de la *Haar Wavelet Transform* del descriptor de color *Scalable Color Descriptor*

El resultat és un conjunt de 16 bins de baixa freqüència i 240 bins d'alta freqüència, els quals tendeixen a tenir valors (positius i negatius) petits. Aquesta propietat es pot explotar de dues formes (dos tipus d'escalabilitat del descriptor):

- Tots o alguns dels bins d'alta freqüència resultants de la HWT poden ser descartats, proporcionant descriptors amb menys nombre de bins, començant per l'original de 256 bins i disminuint a 128, 64, 32 i 16 bins. En aquest últim cas, només es conservarien els coeficients de baixa freqüència. A la **Fig. 1.13 (b)** es pot observar l'**escalabilitat** resultant del SCD.
- El número de bits utilitzats es pot reduir escalant els coeficients a diferents nombres de bits. Aquest tipus d'escalabilitat no s'ha tractat en la implementació del descriptor.

La **Fig. 1.13 (a)** mostra la unitat bàsica de la Transformada *Wavelet Haar* i la **Fig. 1.13 (b)** és un diagrama esquemàtic de la generació del SCD.

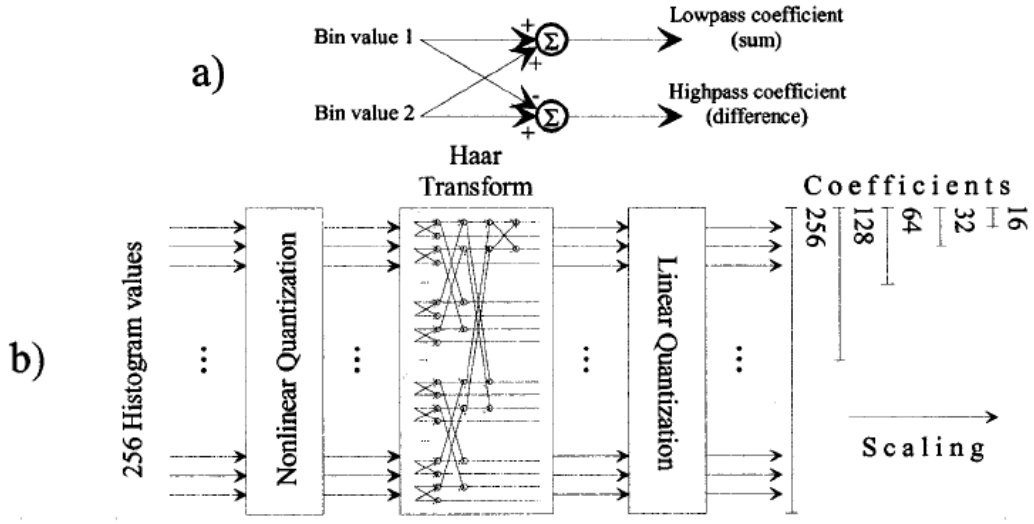


Fig. 1.13 (a) Unitat bàsica de la Transformada *Wavelet Haar*
(b) Diagrama de la generació del SCD

1.6. *Group of Frames Descriptor*

Degut a l'augment del tamany de les bases de dades multimèdia, es fa necessària la representació del contingut de *groups-of-frames* (GoF) o grups de *shots* de vídeo amb descriptors efectius i eficients.

Una vegada són identificats els límits d'un *shot* en una seqüència de vídeo, se sol descriure el contingut visual i de color dels *shots* mitjançant *key frames* (són els representatius en un conjunt de *frames*) i histogrames dels *key frames*, respectivament.

Encara que l'histograma de color dels *key frames* és simple i computacionalment senzill, la descripció de color que proporciona varia significativament en funció del criteri de selecció d'aquests *frames*.

Per a solucionar les variacions en la descripció del color d'un *shot* provocades per les arbitrarietats inherents de la selecció del *key frame*, s'utilitza una aproximació més favorable en la qual es considera el contingut de color de tots els frames d'un *shot* per al càlcul de l'histograma de color.

El *Group-of-Frames/Pictures Descriptor* definit en l'estàndard MPEG-7, és un descriptor de color basat en histograma per a un grup de *frames* o imatges.
Bàsicament, el descriptor GoF és una extensió del *Scalable Color Descriptor*.

Els tres histogrames que defineix s'obtenen acumulant els histogrames de color de múltiples *frames* i representant-los mitjançant el SCD. Aquestes tres formes definides per a calcular els valors l'histograma de color acumulat per un conjunt d'imatges o *frames* d'un vídeo són:

- *Average Histogram*
- *Median Histogram*
- *Intersection Histogram*

El descriptor es pot utilitzar per a diferents aplicacions de recuperació basada en QBE (*query-by-example*), d'**agrupament de *shots*** i de ràpida recerca d'una imatge o d'un vídeo en una base de dades.

1.6.1. *Key frame Histogram*

L'aproximació més senzilla per a descriure el contingut de color d'un GoF és mitjançant l'histograma de color del *frame* representatiu.
Encara que aquesta aproximació és senzilla i, normalment, la menys costosa computacionalment, la descripció de color proporcionada per l'histograma del *key frame*, varia significativament depenent del criteri de selecció del *key frame*.

Un mètode per a seleccionar el *key frame* òptim, per a després obtenir el seu histograma [5], és definint l'**error mitjà** per un frame *l* arbitrari en un GoF donat i el seu respectiu histograma H_l, com:

$$E_{H_l} = \frac{1}{N}\sum_{i=1}^{N}\left\|H_i - H_l\right\| \qquad \textbf{(1.4)}$$

L'histograma que minimitza l'error mitjà, se selecciona com l'histograma del *key frame* òptim.
Encara que l'aproximació que s'obté és consistent, computacionalment és exigent, ja que requereix una recerca exhaustiva sobre tots els *frames* d'un GoF per a determinar quin és el que minimitza l'error mitjà.
Per a determinar l'histograma òptim en un GoF de durada *N*, són necessàries *N·(N-1)* comparacions d'histograma, la qual cosa és costosa en quan a temps de càlcul.

Aquest procediment sempre obté un únic *key frame* per a un GoF; per això, per a obtenir un *key frame* útil convenen més aquells GoFs amb contingut de color més o menys uniforme.

1.6.2. *Average Histogram*

L'*average histogram* (histograma mitjà) és simplement la mitjana dels histogrames de color de tots els *frames* inclosos dins el *shot*.
Es defineix cada bin *j* de l'histograma mitjà pel k-èssim GoF com,

$$AvgHist_k(j) = \frac{1}{N}\sum_{i=b_k}^{e_k} H_i(j), \qquad j = 1,...,B$$

(1.5)

On,

H_i: indica l'histograma del frame i-èssim
b_k: *frame* inicial del GoF_k.
e_k: *frame* final del GoF_k.
N: número de *frames* en el GoF i correspon a $N=(e_k-b_k+1)$

1.6.3. *Median Histogram*

Una altra forma d'obtenir un descriptor de color robust és substituir l'average histogram pel median histogram, el qual pot eliminar eficientment els "outliers" de les dades.

El *median histogram* (histograma mediana) s'obté calculant la mediana de cada bin de l'histograma sobre tots els frames del *shot* i assignant aquest valor al bin resultant de l'histograma.
Cada bin de l'histograma pel k-èssim GoF ve donat per,

$$MedHist_k(j) = median\{H_{b_k}(j), H_{b_k+1}(j),..., H_{e_k-1}(j), H_{e_k}(j)\}$$

(1.6)

Per a calcular-ho, es construeix una llista ascendent de valors d'histograma del *frame* per tot el GoF. El valor central d'aquesta llista s'assigna al valor del bin corresponent. Quan el número de frames, *M*, en el GoF, és parell, el valor central es defineix com el terme mitjà dels dos valors centrals de la llista ordenada.

1.6.4. *Intersection Histogram*

Aquest histograma s'obté calculant el valor mínim de cada bin de l'histograma sobre tots els frames del *shot* i assignant aquest valor al bin resultant de l'histograma.

$$IntHist_k(j) = \min_i \{H_i(j)\} \tag{1.7}$$

Cada valor d'un bin (j) representa el número de píxels d'un color particular que apareix en tots els *frames* del GoF.

A diferència dels histogrames *average* i *median*, proporciona les característiques de color menys poc comunes d'un GoF, enlloc d'una estimació de la distribució de color.
Per això, l'*intersection histogram* és apropiat per a la ràpida **identificació** d'un GoF del qual se'n dóna una imatge.

Per definició tenim,

$$H_i^k(j) \geq IntHist_k(j), \qquad \forall i \in GoF_k, j = 1,...,B \tag{1.8}$$

On H_i^k és l'histograma del *frame i* dins el GoF_k. Aquesta relació és evident ja que l'*intersection histogram* troba el valor mínim del bin i per tant, en un *frame i* qualsevol, aquest valor ha de ser superior o en tot cas igual.

Aquesta propietat pot servir per a determinar el GoF en el qual pertany un *frame*. A partir d'un frame donat *f*, podem trobar els GoFs on no hi és present i descartar-los.

1.7. Algorismes de detecció de *shots*

En aquest apartat s'expliquen els algorismes implementats per a la detecció de *shots*, així com les mesures de similitud aplicades entre els *frames* consecutius de les característiques extretes amb el descriptor SCD per a poder detectar les diferències existents entre *frames*.

El desenvolupament dels algorismes s'ha realitzat, per:

- Mesures de similitud entre *frames* consecutius dels paràmetres estadístics (**μ** i **σ**) del canal HSV.
- Mesures de similitud entre *frames* consecutius dels **bins** extrets del SCD.

1.7.1. Càlcul de mesures de similitud

S'han utilitzat diferents mesures de distància per a, posteriorment, realitzar un contrast (veure apartat 3.1.) entre elles per a veure quina aporta millors resultats en la detecció de *shots*.

- Mesura de distància L_2

$$L_{2\,frame} = \sqrt{\sum_{b=1}^{N} (H_{f+1}[b] - H_f[b])^2} \qquad \textbf{(1.9)}$$

On,

b: índex del bin de l'histograma, b=(1,...,N)
f: frame
Hf: Histograma del frame f

- Mesura de distància χ^2

$$\chi^2_{frame} = \sum_{b=1}^{N} \frac{\left|H_{f+1}[b] - H_f[b]\right|^2}{H_{f+1}[b]} \qquad \textbf{(1.10)}$$

1.7.2. Algorismes de detecció de *Hard Cuts*

Per a la detecció de canvis de càmera es considera una finestra fixa de 21 *frames*.
Tot seguit es calcula la mitjana de (μ_{left} i μ_{right}) i la desviació estàndard (σ_{left} i σ_{right}) de la mesura de similitud dels 10 *frames* de l'esquerra i els 10 de la dreta del frame central de la finestra [6].

El frame en curs serà detectat com a canvi de *shot* si:

1. La mesura de distància del *frame* és la màxima en la finestra.

2. El valor és major que un llindar (*threshold*) adaptatiu T_{cut}.
   On, $T_{cut} = \max(\mu_{left} + \alpha*\sigma_{left} , \mu_{right}+\alpha*\sigma_{right})$, $\alpha=3$.

Aquestes dues condicions asseguren que només els *frames* que es troben dins la finestra especificada, decideixen si hi ha un canvi de *shot* al *frame* central.

Cal comentar, que s'ha descartat utilitzar llindars fixes per a tot el segment de vídeo, ja que presenta dos problemes importants:

1. La distància corresponent a dos canvis de pla diferents pot variar depenent de les característiques del *shot* (moviment, condicions d'il·luminació, etc).

Per tant, és fa molt difícil detectar tots els canvis de càmera si entre uns i altres hi ha molta diferència.

2. No es pot definir un sol llindar fix per qualsevol gènere de vídeo, ja que el contingut d'un partit de bàsquet és molt diferent a el d'un programa informatiu. Per això s'hauria de realitzar la tasca d'ajustar el llindar per a cada gènere de vídeo diferent.

A la **Fig. 1.14** es presenta el diagrama de flux de l'algorisme implementat:

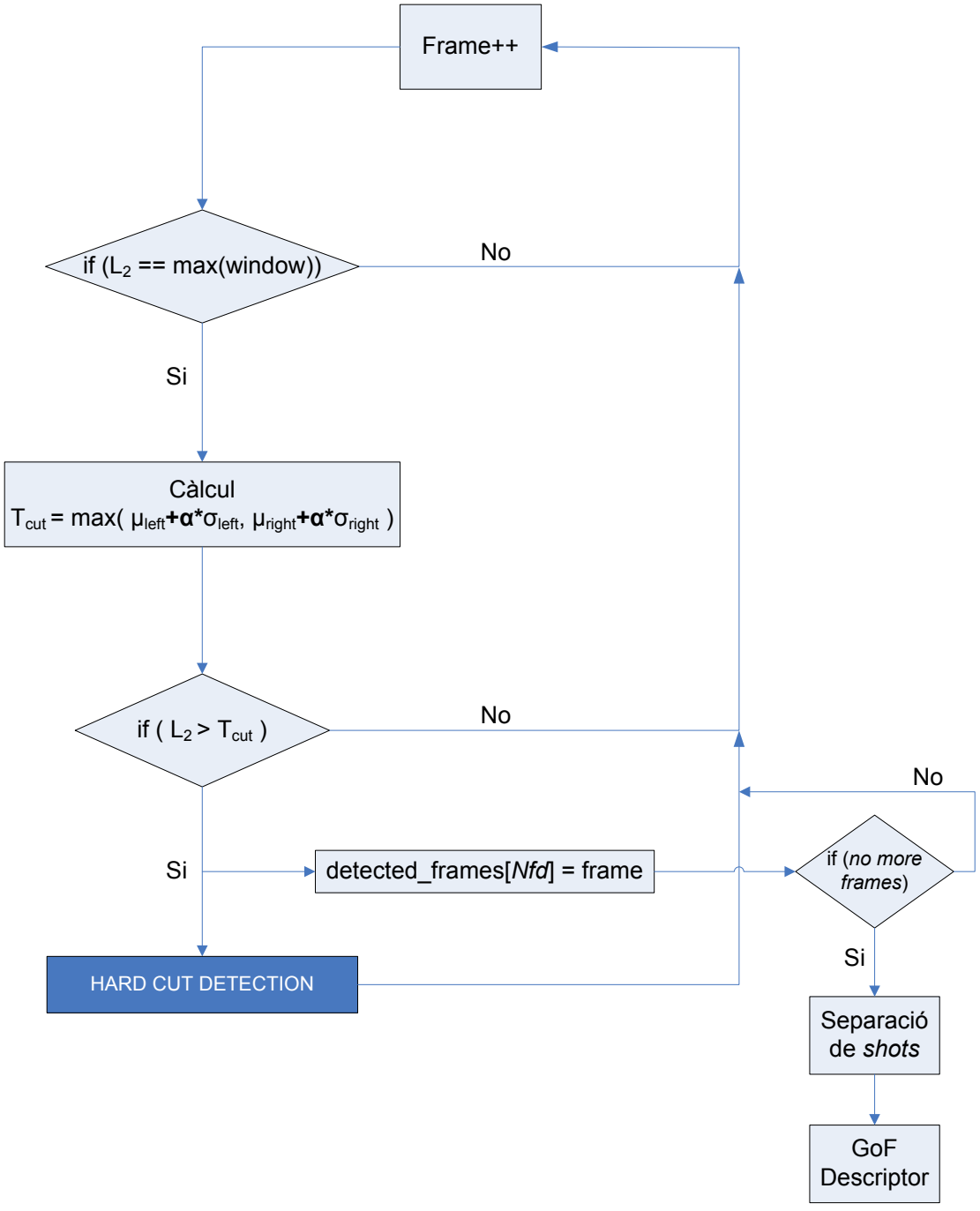


Fig. 1.14 Diagrama de flux de l'algorisme de detecció de *Hard Cuts*

Una vegada s'ha conclòs la detecció dels *hard cuts* en el segments de vídeo, es descriu el contingut de cadascun dels *shots* mitjançant el descriptor GoF.

1.7.3. Algorismes de detecció de *Soft Cuts*

Per a la detecció de transicions graduals es considera una finestra de 21 *frames*, 10 *frames* a l'esquerra i 10 a la dreta del *frame* central de la finestra, el qual és testejat per l'algorisme.

S'aplica un filtre mediana a les mesures de distància entre *frames* consecutius per a eliminar els pics i les fluctuacions, aconseguint suavitzar els valors d'aquestes diferències entre *frames*.

Tot seguit es calcula la mitjana de (μ_{left} i μ_{right}) i la desviació estàndard (σ_{left} i σ_{right}) de la mesura de similitud dels *frames* de la finestra.

Es defineixen dos llindars adaptatius:
$T_{high} = \max(\mu_{left} + \alpha_{high}*\sigma_{left} , \mu_{right}+\alpha_{high}*\sigma_{right})$
$T_{low} = \max(\mu_{left} + \alpha_{low}*\sigma_{left} , \mu_{right}+\alpha_{low}*\sigma_{right})$

Es marca l'inici d'una transició gradual si la mesura de distància del *frame* en curs es troba entre els dos llindars definits.
A partir d'aquest inici es comença a fer una acumulació de la comparació entre *frames*, fins que la diferència entre *frames* consecutius descendeixi fins a ser menor que el llindar inferior i l'acumulació computada sigui major que el llindar superior [7] [8].

A la **Fig. 1.15** es presenta el diagrama de flux de l'algorisme implementat:

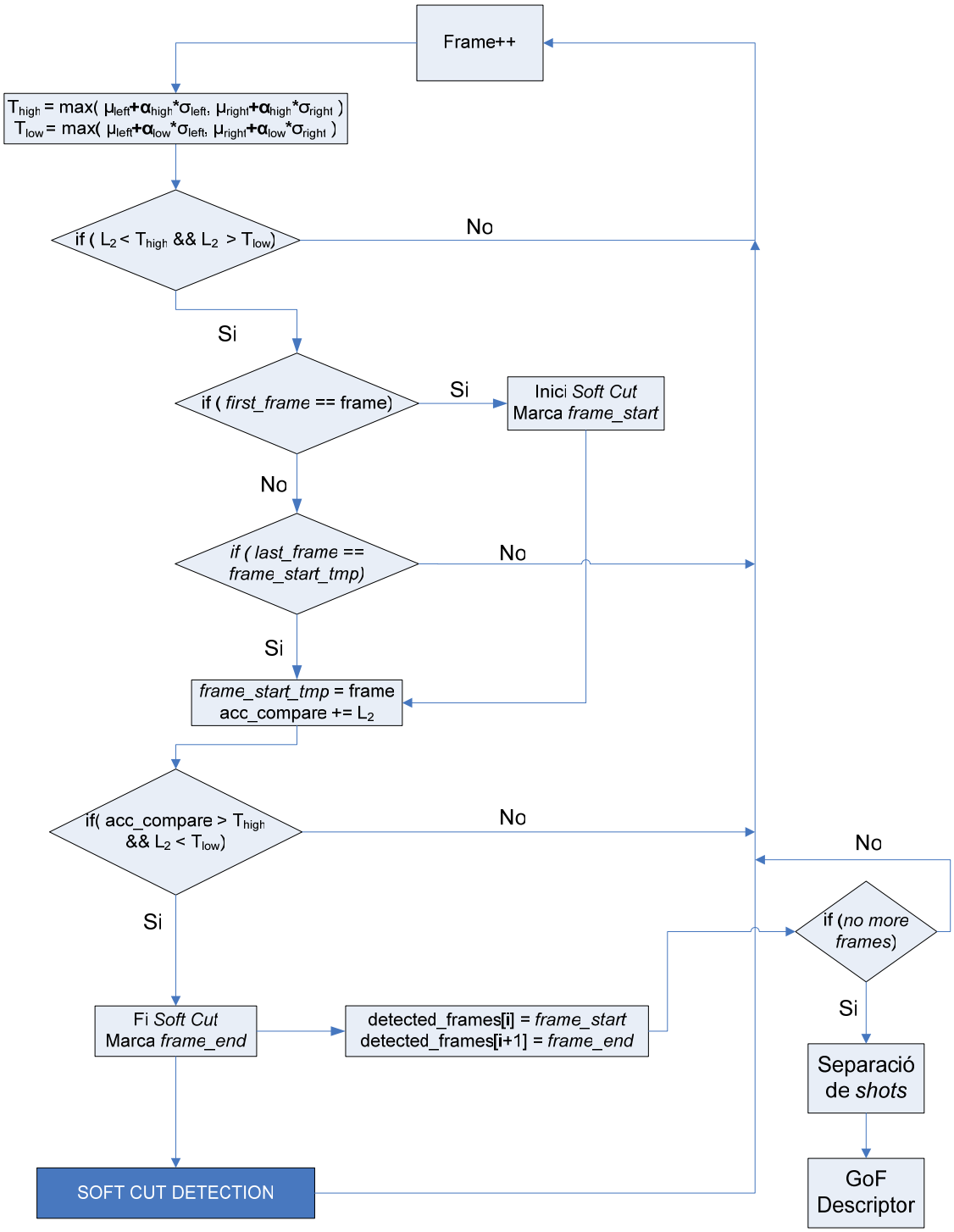


Fig. 1.15 Diagrama de flux de l'algorisme de detecció de *Soft Cuts*

Una vegada s'ha conclòs la detecció dels *soft cuts* en el segments de vídeo, es descriu el contingut de cadascun dels *shots* mitjançant el descriptor GoF.

CAPÍTOL 2. SEGMENTACIÓ I DESCRIPCIÓ DE MATERIAL AUDIOVISUAL

Per a poder implementar i posteriorment comprovar els resultats d'algorismes de **segmentació** i **classificació** de segments audiovisuals és necessari tenir una base de dades audiovisual degudament segmentada i etiquetada manualment.

Si la utilitat de la base de dades està destinada a classificar el contingut audiovisual que contenen és necessari que aquesta disposi d'un ampli etiquetatge del seu contingut. Si, per altra banda, la utilitat és comprovar algorismes de segmentació automàtica de vídeo no és necessari disposar d'un etiquetatge tant extens.

En aquest treball, el material audiovisual és útil per a comprovar i analitzar si els algorismes implementats funcionen.

El material consisteix en una hora de material AV de la *Televisió de Catalunya* originalment enregistrat en MPEG-2 *High Quality* (720x576 @ 10Mbps).

Tipus de material:

- Esports (*Sports*)
- Notícies (*News*)
- Publicitat (*Commercials*)
- Telenovela (*Soap Opera*)
- Documental (*Documental*)

El vídeo original ha estat dividit en varis vídeos de durada més curta. Aquests han estat segmentats manualment en **escenes** d'alt nivell semàntic, és a dir, descrivint el contingut audiovisual en el seu nivell més específic.

La **Fig. 2.1** il·lustra una sèrie de canvis de pla dins una escena d'una telenovela, corresponent a un diàleg entre dues persones. Aquí podem veure el nivell més específic.

Fig. 2.1 Exemple d'una escena d'una seqüència de vídeo amb 6 *shots*

D'aquesta segmentació s'obtenen arxius segmentats dels vídeos i etiquetes en corresponents fitxers adjunts (.txt), els quals contenen la descripció corresponent de cada arxiu.

Al final del capítol, es descriu la forma en la qual s'han estructurat les dades. En un DVD s'han guardat els vídeos originals i els vídeos segmentats amb les seves corresponents etiquetes.
Per altra banda, en un altre DVD, s'han guardat fitxers de text de característiques tant d'àudio com de vídeo per a cada segment de vídeo. Això s'ha realitzat gràcies a un software ja existent.

2.1. Descripció de les etiquetes

Els fitxers corresponents als vídeos que s'han segmentat han estat etiquetats de la següent forma:

- Nom del fitxer de vídeo complet
- *Frame Rate* (*Frames* per segon)
- *Time Code* IN & OUT: posició inicial i final del segment de vídeo dins el fitxer original.
- Gènere general
    - Esports (*sports*)
    - Notícies (*news*)
    - Publicitat (*commercials*)
    - Telenovela (*soap opera*)
    - Documental
    - Videoclips
    - Dibuixos animats (*cartoons*)
- Sub-classificació (depenent del gènere general):
    - esports: futbol, bàsquet, corner, gol, penal, tir lliure...
    - notícies: política, socials, temps, esports, economia...
    - publicitat: marca, companyia...
    - telenovela
    - documental: animals, ciències naturals, històric...
- Persones: nombre de persones que apareixen en l'escena
- Parla (*Speech: yes / no*)

- Tipus de parla (*Speech Type: speaker / reporter / both / none*)
- Música (*Music: yes / no*)
- Tipus de música (*Music Type: rock / classic / others / none*)

Podem observar un exemple d'etiqueta a la **Fig. 2.2**. Els fitxers (.txt) corresponents a cada segment de vídeo tenen aquest format.

```
Original_Video VARIS_01
Frame_Rate 25
TC_IN 00;03;25;02
TC_OUT 00;03;54;20
Genre documental
Classification view
Persons 1
Speech yes
Speech_Type reporter
Music yes
Music_Type others
```

Fig. 2.2 Exemple d'etiqueta

2.2. Procés de segmentació

Per a segmentar manualment l'arxiu de vídeo s'ha utilitzat el software *Adobe Premiere 6.5*.
Primerament, s'ha de convertir l'arxiu original (format VOB) ja que l' *Adobe Premiere 6.5*. no accepta el format; per tant, abans de segmentar-ho, és necessari convertir-ho a format MPEG, format llegible pel *Premiere*.

El procés de segmentació es podria definir de la següent manera:

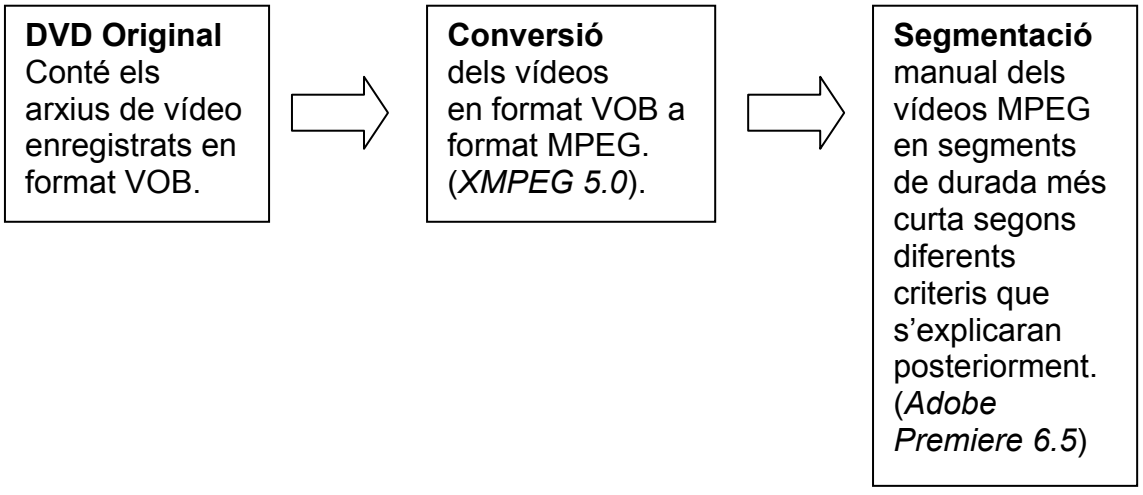


Fig. 2.3 Procés de segmentació del contingut audiovisual

2.2.1. Conversió de format VOB a MPEG

El software utilitzat per a la conversió dels vídeos originals és el *XMPEG 5.0* (http://www.xmpeg.net/). Aquest programa s'ha escollit perquè és gratuït i per el gran ventall de possibilitats a l'hora de realitzar la conversió, en gran part perquè permet mantenir la màxima qualitat que oferia el vídeo original.

La **Fig. 2.3** mostra el format de sortida que s'ha escollit:

MPEG-2	
Resolution:	720x576
Aspect Ratio:	4:3
Frame Rate:	25fps - PAL (625/50)
Video Bit Rate:	6Mbps (constant)
Audio Bit Rate:	384Kbps

Fig. 2.3 Format de sortida

S'han escollit aquestes taxes de bit per a què la qualitat del material sigui acceptable i molt fidel a la original.

2.2.2. Segmentació dels MPEG

La segmentació dels vídeos convertits a MPEG es realitza mitjançant el software *Adobe Premiere 6.5* amb un codificador addicional, el *LSX-MPEG Suite 2.0 - Premiere Plug in*, el qual ens permet manipular més paràmetres que els que ens permet el codificador per defecte del *Premiere*.

Quan a metodologia de segmentació es pot dir que els segments s'han separat, a grans trets, seguint criteris de **canvis en l'àudio** i **canvis d'escena**.

A la **Fig. 2.4** es poden veure els paràmetres utilitzats en la segmentació:

Export Movie Settings:

*File Type: LSX **MPEG-2***
Range: In to Out
Export Video & Audio

Video Settings:
Compressor: LSX-MPEG Video
Frame Size: 720x576
Frame Rate: 25 fps
Pixel Aspect Ratio: D1/DV PAL (1.067)

Audio Settings:
Audio Rate: 48 KHz
Audio Format: 16 Bit - Stereo
Audio Interleave: None
Compressor: LSX-MPEG Layer 2 Audio

Rendering options:
Fields: Upper Field First

LSX-MPEG Settings:

Video Stream Data Rate: 6 Mbps (constant)
Audio Stream Bit Rate: 384Kbps
Aspect Ratio: 4:3 PAL
Main Profile & Main Level

Frame Sequence: I BB P BB P BB P BB I…
(3 P frames & 2 B frames)

Fig. 2.4 Paràmetres de la segmentació MPEG

2.2.3. Conversió als formats AVI i WAV

També s'ha procedit a convertir els segments als formats d'àudio WAV i de vídeo AVI (**Fig. 2.5**) i per a poder després poder introduir-los a l'entrada de l'aplicació implementada per a la detecció de *shots* (només AVI) i també, per a poder extreure a fitxers de text les característiques d'àudio i vídeo de cada vídeo amb el software ja existent.

Export Audio Settings:

*File Type: **Windows Waveform***
Range: In to Out

Audio Settings:
Audio Rate: 48 KHz
Audio Format: 16 Bit - Mono
Compressor: Uncompressed

Rendering options:
Fields: Lower Field First

Export Movie Settings:

File Type: ***Microsoft AVI***
Range: In to Out
Export Video & Audio

Video Settings:
Compressor: Ms MPEG4 Video Codec v2
Frame Size: 352x288
Frame Rate: 25 fps
Pixel Aspect Ratio: D1/DV PAL (1.067)

Audio Settings:
Audio Rate: 48 KHz
Audio Format: 16 Bit - Stereo
Audio Interleave: None
Compressor: Uncompressed

Fig. 2.5 Paràmetres de la segmentació WAV i AVI

2.3. Estructuració de les dades

Per a concloure l'etiquetatge del contingut audiovisual, és necessari mantenir una estructura del material. El contingut AV es divideix en dos DVD de la següent forma:

Contingut del DVD-1:

Podem trobar les dades originals convertides:

VARIS_01.mpg	→	Documental d'animals	(7:30)
VARIS_02.mpg	→	Programa d'entrevistes	(4:11)
VARIS_03.mpg	→	Telenovela	(3:15)
VARIS_04.mpg	→	Publicitat i Documental	(9:22)
VARIS_05.mpg	→	Esports (Bàsquet)	(5:57)
VARIS_06_1.mpg	→	Notícies	(6:03)
VARIS_06_2.mpg	→	Notícies i Publicitat	(8:45)
VARIS_06_3.mpg	→	Notícies i Publicitat	(7:18)
		Total	(52:21)

Conté un directori per a cada fitxer MPEG original, en el qual podem trobar:

- Segments de vídeo en format AVI (352x288) → (*VARIS_0X_XXX_s.avi*)
- Fitxer d'àudio WAV corresponent al segment de vídeo AVI respectiu → (*VARIS_0X_XXX_s.wav*)
- Fitxer TXT amb totes les característiques d'àudio del segment → (*VARIS_0X_XXX_s_AudioFeatures.txt*)
- Fitxer TXT amb totes les característiques de vídeo del segment → (*VARIS_0X_XXX_s_VideoFeatures.txt*)

- Fitxer TXT amb totes les característiques audiovisuals del segment *(VARIS_0X_XXX_s_AVFeatures.txt)*

Tot seguit, es pot veure el format dels directoris:

VARIS_01\
VARIS_01_001_s.avi
VARIS_01_001_s.wav
VARIS_01_001_s_AudioFeatures.txt
VARIS_01_001_s_VideoFeatures.txt
VARIS_01_001_s_AVFeatures.txt
…
VARIS_01_018_s.avi
VARIS_01_018_s.wav
VARIS_01_018_s_AudioFeatures.txt
VARIS_01_018_s_VideoFeatures.txt
VARIS_01_018_s_AVFeatures.txt

VARIS_02\
VARIS_02_001_s.avi
VARIS_02_001_s.wav
VARIS_02_001_s_AudioFeatures.txt
VARIS_02_001_s_VideoFeatures.txt
VARIS_02_001_s_AVFeatures.txt
…
VARIS_02_010_s.avi
VARIS_02_010_s.wav
VARIS_02_010_s_AudioFeatures.txt
VARIS_02_010_s_VideoFeatures.txt
VARIS_02_010_s_AVFeatures.txt

VARIS_03\
VARIS_03_001_s.avi
VARIS_03_001_s.wav
VARIS_03_001_s_AudioFeatures.txt
VARIS_03_001_s_VideoFeatures.txt
VARIS_03_001_s_AVFeatures.txt
…
VARIS_03_005_s.avi
VARIS_03_005_s.wav
VARIS_03_005_s_AudioFeatures.txt
VARIS_03_005_s_VideoFeatures.txt
VARIS_03_005_s_AVFeatures.txt

VARIS_04\
VARIS_04_001_s.avi
VARIS_04_001_s.wav
VARIS_04_001_s_AudioFeatures.txt
VARIS_04_001_s_VideoFeatures.txt
VARIS_04_001_s_AVFeatures.txt
…
VARIS_04_020_s.avi
VARIS_04_020_s.wav
VARIS_04_020_s_AudioFeatures.txt
VARIS_04_020_s_VideoFeatures.txt
VARIS_04_020_s_AVFeatures.txt

VARIS_05\
VARIS_05_001_s.avi
VARIS_05_001_s.wav
VARIS_05_001_s_AudioFeatures.txt
VARIS_05_001_s_VideoFeatures.txt
VARIS_05_001_s_AVFeatures.txt
…
VARIS_05_031_s.avi
VARIS_05_031_s.wav
VARIS_05_031_s_AudioFeatures.txt
VARIS_05_031_s_VideoFeatures.txt
VARIS_05_031_s_AVFeatures.txt

VARIS_06_1\
VARIS_06_1_001_s.avi
VARIS_06_1_001_s.wav
VARIS_06_1_001_s_AudioFeatures.txt
VARIS_06_1_001_s_VideoFeatures.txt
VARIS_06_1_001_s_AVFeatures.txt
…
VARIS_06_1_022_s.avi
VARIS_06_1_022_s.wav
VARIS_06_1_022_s_AudioFeatures.txt
VARIS_06_1_022_s_VideoFeatures.txt
VARIS_06_1_022_s_AVFeatures.txt

VARIS_06_2\
VARIS_06_2_001_s.avi
VARIS_06_2_001_s.wav
VARIS_06_2_001_s_AudioFeatures.txt
VARIS_06_2_001_s_VideoFeatures.txt
VARIS_06_2_001_s_AVFeatures.txt
…
VARIS_06_2_024_s.avi
VARIS_06_2_024_s.wav
VARIS_06_2_024_s_AudioFeatures.txt
VARIS_06_2_024_s_VideoFeatures.txt
VARIS_06_2_024_s_AVFeatures.txt

VARIS_06_3\
VARIS_06_3_001_s.avi
VARIS_06_3_001_s.wav
VARIS_06_3_001_s_AudioFeatures.txt
VARIS_06_3_001_s_VideoFeatures.txt
VARIS_06_3_001_s_AVFeatures.txt
…
VARIS_06_3_028_s.avi
VARIS_06_3_028_s.wav
VARIS_06_3_028_s_AudioFeatures.txt
VARIS_06_3_028_s_VideoFeatures.txt
VARIS_06_3_028_s_AVFeatures.txt

Contingut del DVD-2:

- Document on es descriu tot el procés → *AV-Material-VARIS.doc*
- Conté un directori per a cada fitxer MPEG original, en el qual podem trobar:
  - Segments de vídeo en format DVD de (720x576) píxels → (*VARIS_0X_XXX.mpg*)
  - Segments de vídeo en format CIF (*Common Intermediate Format*) corresponent a (352x288) píxels → (*VARIS_0X_XXX_s.mpg*)
  - Etiquetes corresponents a cada segment → (*VARIS_0X_0XX.txt*)
  - Fitxer TXT amb totes les etiquetes → (*VARIS_0X.txt*)

Tot seguit, es pot veure el format dels directoris:

VARIS_0X\
VARIS_0X.txt
VARIS_0X_00X.mpg
VARIS_0X_00X_s.mpg
VARIS_0X_00X.txt
…
VARIS_0X_0XX.mpg
VARIS_0X_0XX_s.mpg
VARIS_0X_0XX.txt

CAPÍTOL 3. RESULTATS

Per a avaluar els algorismes implementats, és necessari disposar d'una base de dades audiovisual.
Un cop s'ha elaborat la tasca de segmentació i descripció manual d'aquest material, com s'ha explicat en el capítol anterior, es poden comparar els resultats obtinguts amb el mètode automàtic respecte els que s'ha etiquetat manualment.

En l'apartat 1.7. s'expliquen els algorismes implementats per a la detecció de per a *hard cuts* i *soft cuts*. A causa del temps limitat per a la realització del treball, només s'han pogut comprovar els resultats per a la detecció de *hard cuts*, ja que, a pesar d'haver implementat l'algorisme per a la detecció de *soft cuts*, no s'ha tingut temps d'ajustar els paràmetres (α) de forma adequada.

3.1. Contingut visual

El problema dels mètodes de detecció automàtica de *shots* és que per a comprovar els resultats, s'ha de disposar d'una etiquetatge dels *shots* presents en els segments de vídeo, ja que no existeix cap estàndard de vídeo relacionat amb la detecció-representació de *shots.*

La **Taula 3.1** mostra el contingut i l'extensió del material audiovisual utilitzat (veure Capítol 2 per a més detalls sobre el contingut).
Sense aquesta informació no es mostraria clarament el funcionament de l'algorisme. Per a obtenir resultats precisos és important que hi hagi tot tipus de gèneres de vídeo, amb característiques molt diferenciades (molt de moviment en un partit de bàsquet, moviment gairebé nul en la presentació d'un programa informatiu, etc).

Taula 3.1. Contingut del material audiovisual utilitzat

Tipus de vídeo	Durada	# Shots (Hard/Soft Cuts)
Documental d'animals	7' 30"	55 / 2
Programa d'entrevistes	4' 11"	39 / 22
Telenovela	3' 15"	22 / 0
Publicitat i documental	9' 22"	99 / 10
Esport (Bàsquet)	5' 57"	24 / 10
Notícies i publicitat	4' 35"	49 / 6
TOTAL	**34' 50"**	**288 / 50**

3.2. Paràmetres d'avaluació

Amb l'objectiu d'avaluar els algorismes de detecció de *shots* implementats, els resultats obtinguts es comparen amb un fitxer de referència. Aquest fitxer consta de fulles de càlcul d'*Excel* per a cada gènere audiovisual existent.
La **Taula 3.2** és un exemple que il·lustra el format utilitzat en el fitxer de referència, on es poden observar els diferents camps etiquetats per a cada segment de vídeo.

Taula 3.2 Exemple del fitxer de referència

Segment	Sub-segment	Frames	Tipus Transició	Detecció
VARIS_01				
	V_01_005			
		40-41	*Hard Cut*	*Detected*
		222-223	*Hard Cut*	*Detected*
		871-872	*Hard Cut*	*Detected*
		1077-1078	*Hard Cut*	*Detected*
		1181-1182	*Hard Cut*	*Detected*
		1275-1276	*Hard Cut*	*Detected*
		1404-1405	*Hard Cut*	*Detected*
		1644-1645	*Hard Cut*	*Detected*
	V_01_006			
		147-167	*Soft Cut*	*Missed-detected*
		157-158	*Hard Cut*	*False-detected*

Per a l'avaluació dels resultats obtinguts en el sistema implementat es consideren una sèrie de paràmetres:

- N_d (*detected*):nombre de *shots* detectats correctament.
- N_m (*missed*): nombre de *shots* no detectats (perduts).
- N_f (*false detected*): nombre de *shots* detectats incorrectament.
- N_t (=Nd+Nm): nombre total de *shots* presents en els segments de vídeo.

A partir d'aquests paràmetres s'utilitzen dues mesures d'avaluació de l'algorisme, anomenades *Recall* i *Precision*.

$$\textbf{\textit{RECALL}} = \frac{N_t - N_m}{N_t} = \frac{N_d}{N_t} \qquad \textbf{(3.1)}$$

$$\textbf{\textit{PRECISION}} = \frac{N_t - N_m}{(N_t - N_m) + N_f} = \frac{N_d}{N_d + N_f} \qquad \textbf{(3.2)}$$

Recall mesura el percentatge de deteccions de *shots*. A la fórmula (3.1) s'observa que només es tenen en compte les deteccions que no s'han detectat (N_m), no té en compte si les deteccions són correctes o incorrectes.

Per això necessitem una altra mesura, *Precision*, per a comprovar l'eficiència dels resultats. La mesura *Precision* és el percentatge que mostra com d'acurat és l'algorisme a l'hora de no detectar *shots* incorrectes (N_f).
Com s'ha explicat en l'apartat 1.7., s'han utilitzat dues mesures de distància diferents per a la detecció de canvis entre *frames*, per al posterior descart d'una d'elles, observant els resultats que s'obtenen quan s'analitzen una sèrie de vídeos de contingut variat. La **Fig. 3.1** mostra el resultat de la comparació de les mesures L_2 i χ^2 (a la gràfica correspon a *Chi Square*) corresponents a paràmetres estadístics.

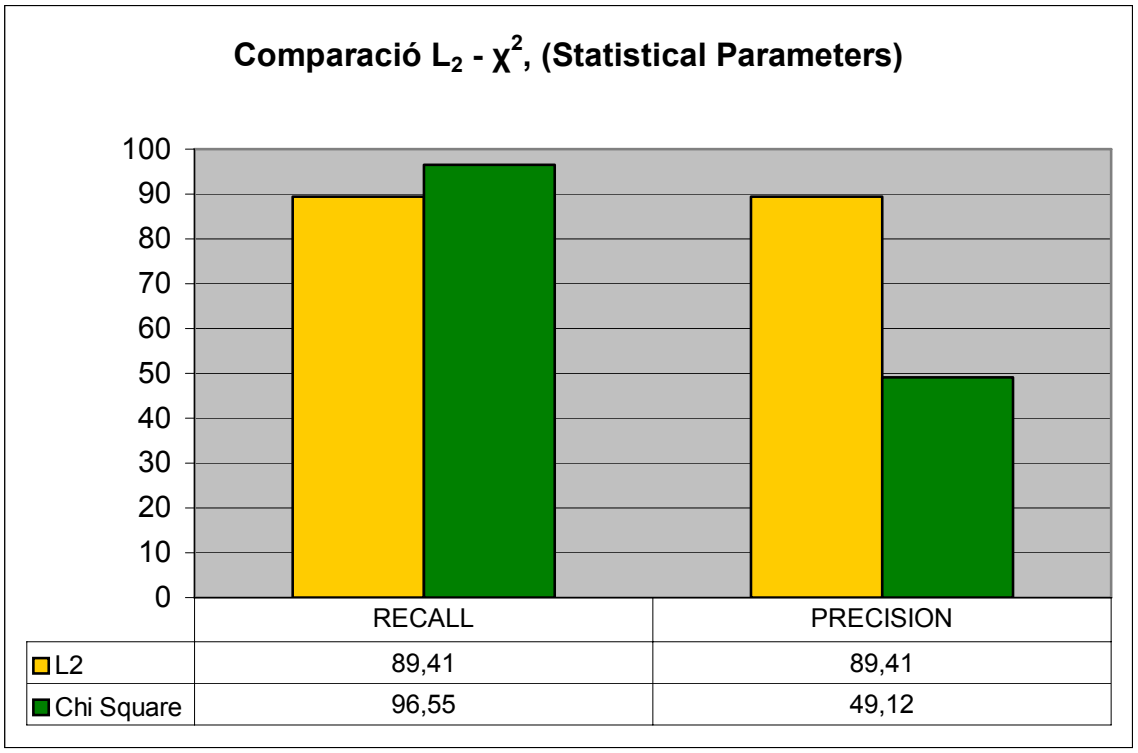


Fig. 3.1 Resultats de la comparació de la detecció de *hard cuts* en una sèrie de vídeos pertanyents a diversos gèneres

Per la mesura L_2 tenim *Recall* i *Precision* gairebé al 90%, en canvi per a la mesura χ^2, *Recall* és més gran (97%) ja que es detecten més *hard cuts* però és molt poc precís (*Precision* = 49%), és a dir, menys de la meitat de deteccions que fa són correctes.
Veient aquests resultats, es descarta la mesura de distància χ^2 i es procedirà a avaluar els algorismes només amb la mesura de distància L_2, tant pels paràmetres estadístics del canal HSV com pels bins de l'histograma extret pel SCD.

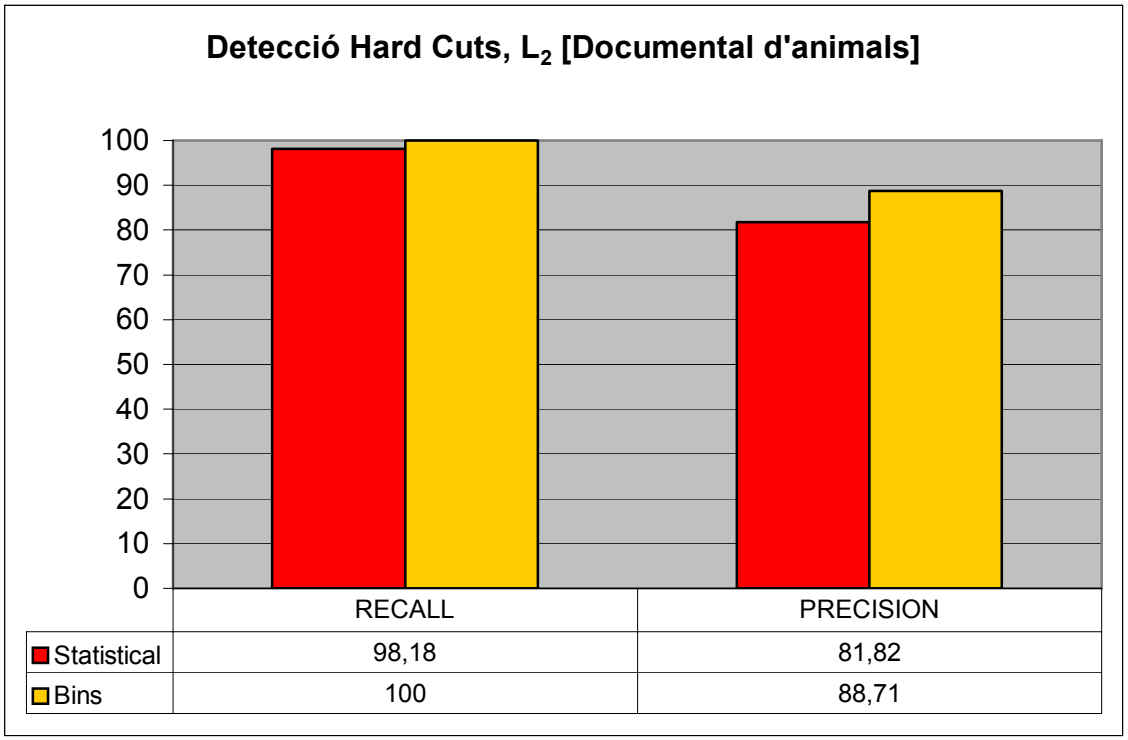
3.3. Resultats experimentals

Es presenten els resultats experimentals obtinguts a partir de l'avaluació dels algorismes. Aquesta avaluació es realitza de dues formes diferents:

- **Statistical**: A partir de la mesura de distància L_2 entre *frames* consecutius per als paràmetres estadístics μ i σ del canal HSV.
- **Bins**: A partir de la mesura de distància L_2 entre *frames* consecutius per als bins de l'histograma extret pel SCD.

3.3.1. Documental d'animals (*VARIS_01*)

La **Fig. 3.2** mostra el resultat de la detecció de *hard cuts* en documentals d'animals:



	RECALL	PRECISION
Statistical	98,18	81,82
Bins	100	88,71

Fig. 3.2 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere de documental (*VARIS_01*)

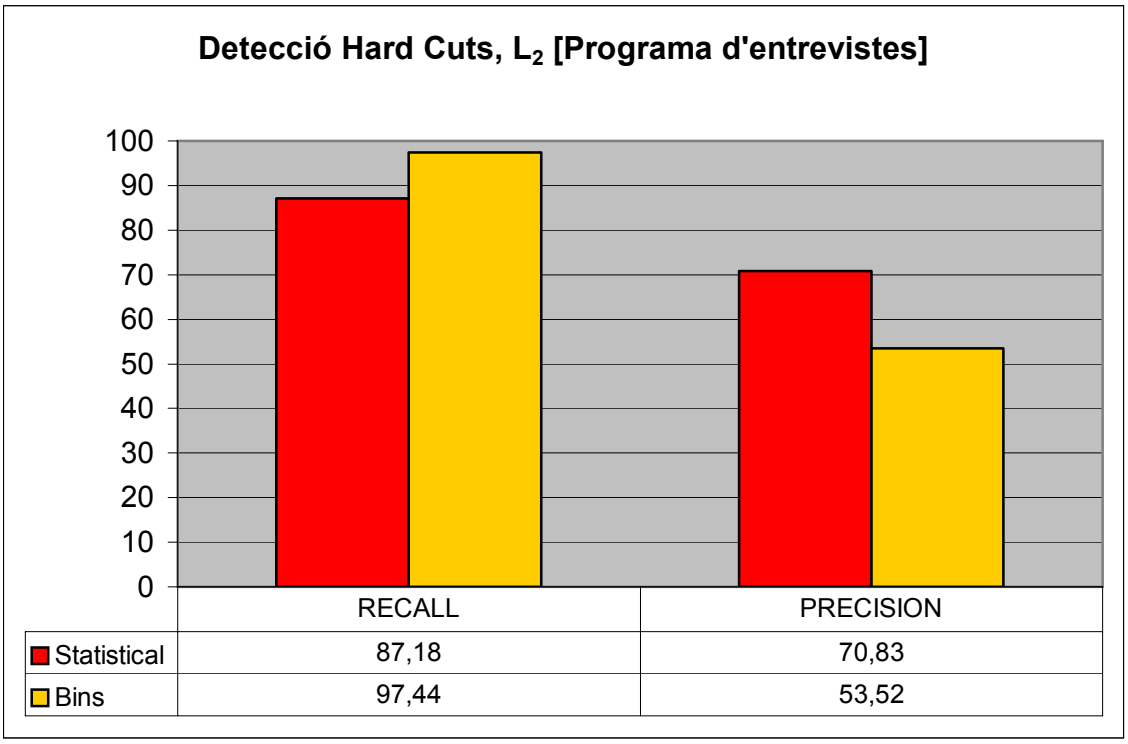
Per a **Statistical**, es detecten gairebé la totalitat (98%) dels *hard cuts* (gairebé no hi ha deteccions perdudes) i s'obté un alt percentatge (82%) de precisió en les deteccions.
Per a **Bins**, es detecten el 100% de *hard cuts* (deteccions perdudes = 0), major que per a Statistical, mentre que la precisió també és més gran (89%) que la obtinguda per a Statistical.

Aquests percentatges tant elevats apareixen per la naturalesa del vídeo d'un documental, el qual consta, generalment, de poc moviment, de canvis de pla considerablement bruscs i de poques transicions graduals (2 per ser exactes). Per això el *Recall* frega la perfecció i hi ha un percentatge bastant alt de *Precision*, el qual reflexa les deteccions incorrectes.

3.3.2. Programa d'entrevistes (*VARIS_02*)

En un programa on s'han editat manualment transicions graduals amb efectes especials entre cada *shot*, la precisió en les deteccions serà considerablement més baixa que, per exemple, el gènere de vídeo corresponent al documental comentat al subapartat anterior on no hi ha amb prou feines edició addicional entre canvis de *shots*.

La **Fig. 3.3** mostra el resultat de la detecció de *hard cuts* en un programa d'entrevistes:



	RECALL	PRECISION
Statistical	87,18	70,83
Bins	97,44	53,52

Fig. 3.3 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere de programa d'entrevistes (*VARIS_02*)

Per a **Statistical**, es detecten un alt nombre de *hard cuts*, això queda reflectit a Recall (87%). S'observa que la precisió no és baixa (71%) però tampoc és alta, això és deu al fet que es fan deteccions errònies quan hi ha efectes especials entre els *shots*.

Per a **Bins**, es detecten gairebé la totalitat (97%) dels *hard cuts* (gairebé no hi ha deteccions perdudes) però la precisió és molt baixa (54%) degut a l'efecte explicat.

El fet d'haver-hi, relativament, molts efectes especials provoca una sobredetecció desmesurada, sobretot en el cas de **Bins**, que es reflecteix en una *Precision* molt baixa.

3.3.3. Telenovela (*VARIS_03*)

En aquesta telenovela en concret, no hi ha transicions graduals, la qual cosa provoca que pràcticament no es realitzin deteccions errònies.

La **Fig. 3.4** mostra el resultat de la detecció de *hard cuts* en una telenovela:

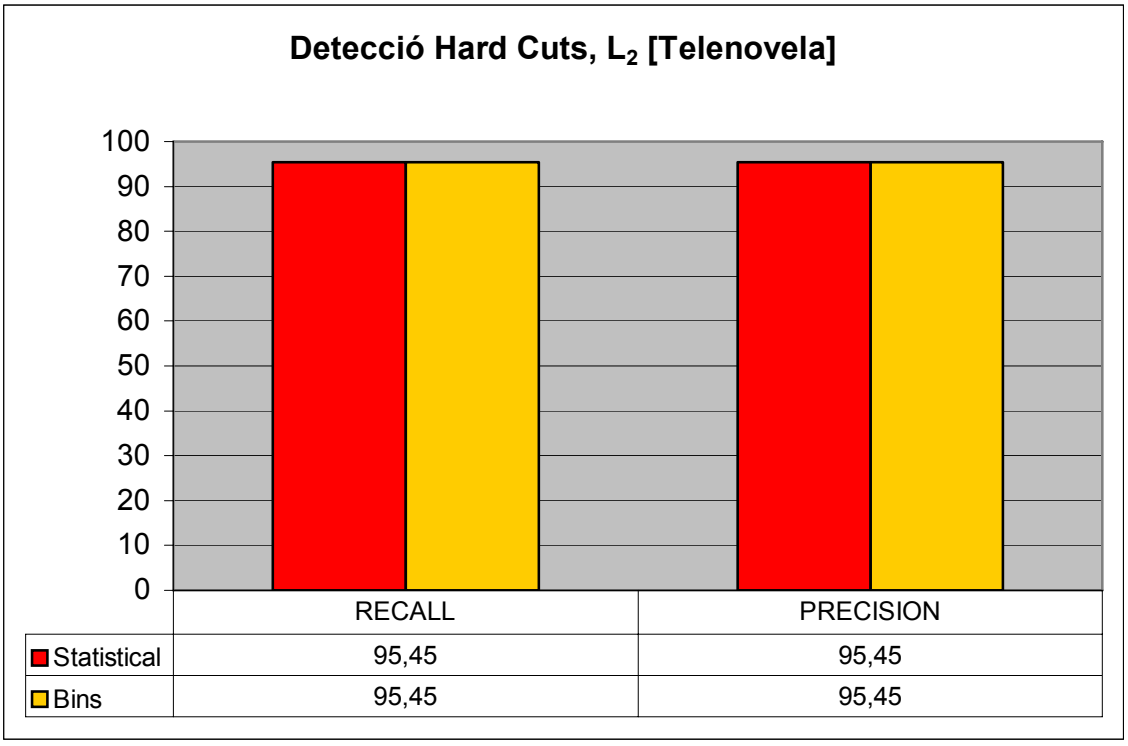


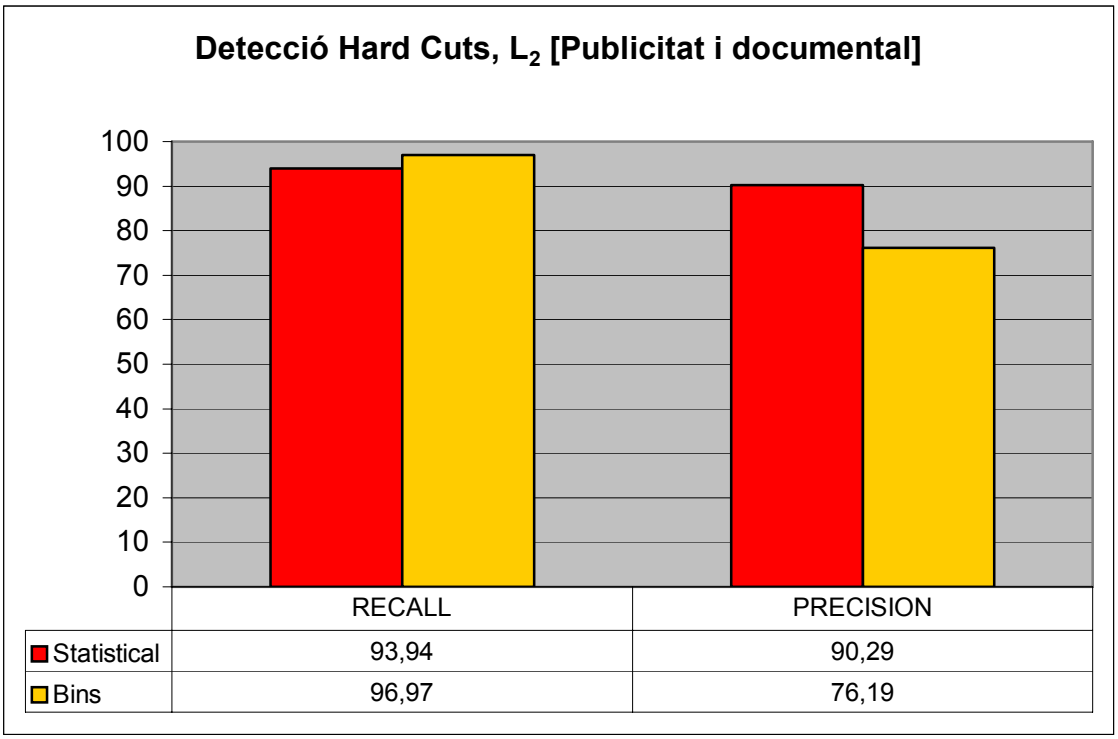
Fig. 3.4 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere de telenovela (*VARIS_03*)

S'obtenen molt bons resultats tant per **Statistical** com per **Bins**. Recall i Precision presenten uns valors percentuals molt elevats (95%) i, curiosament, són iguals, també degut a què hi ha pocs *shots* per a detectar en aquest cas, i és més fàcil que els resultats coincideixin.

3.3.4. Publicitat i documental (*VARIS_04*)

Els anuncis publicitaris, igual que en el programa d'entrevistes comentat al subapartat 3.3.2., han estat objecte d'un procés d'edició i, per tant, poden presentar transicions graduals amb efectes especials entre *shots*, la qual cosa pot provocar deteccions errònies.

La **Fig. 3.5** mostra el resultat de la detecció de *hard cuts* en anuncis publicitaris entre un documental:



	RECALL	PRECISION
Statistical	93,94	90,29
Bins	96,97	76,19

Fig. 3.5 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere de publicitat i documental (*VARIS_04*)

Per a **Statistical**, es detecten un alt nombre de *hard cuts*, això queda reflectit a Recall (94%). S'observa que la precisió en les deteccions és considerablement alta (90%).
En canvi, per a **Bins**, es detecten gairebé la totalitat (97%) dels *hard cuts* (gairebé no hi ha deteccions perdudes) però la precisió és més baixa (76%) que per a Statistical, això és degut a què les mesures de similitud entre bins d'histogrames són més sensibles a la detecció de *hard cuts*.
Això provoca que on hi ha una transició gradual, sovint es detecti un canvi de càmera, augmentant el nombre de deteccions errònies, la qual cosa es reflecteix en un descens en la precisió.

3.3.5. Esports (*VARIS_05*)

La **Fig. 3.6** mostra el resultat de la detecció de *hard cuts* en un partit de bàsquet:

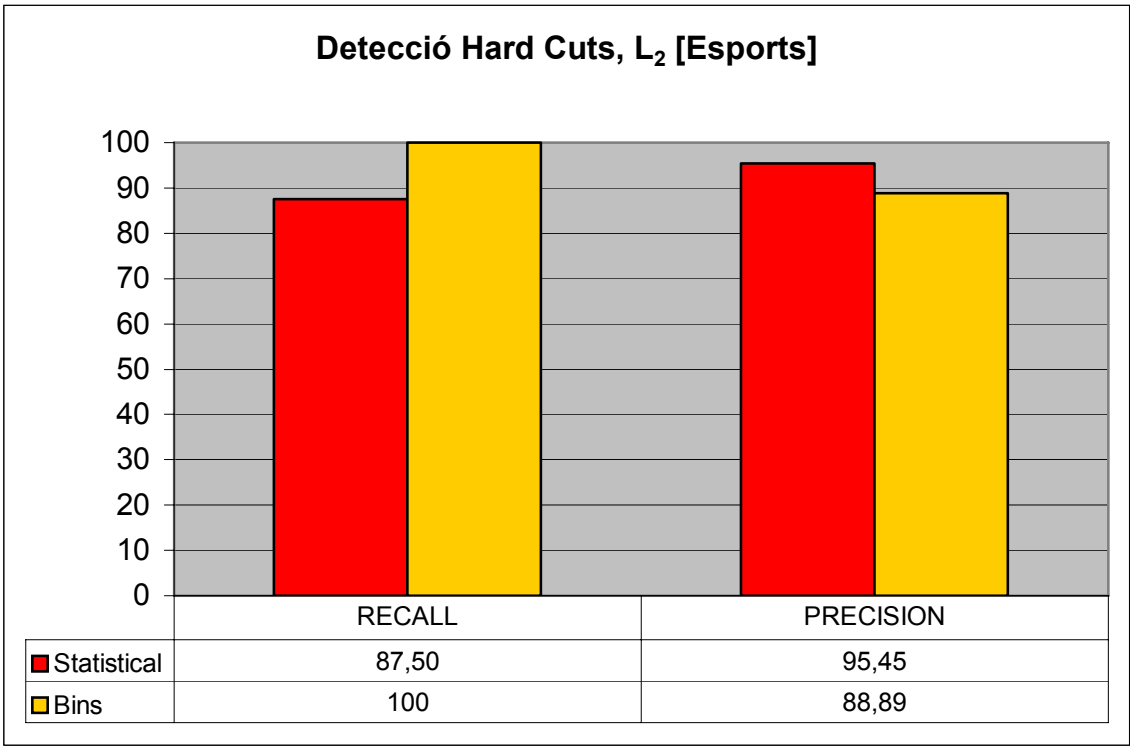


Fig. 3.6 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere d'esports (*VARIS_05*)

Per a **Statistical**, es detecten un alt nombre de *hard cuts*, això queda reflectit a Recall (88%) i s'obté un altíssim percentatge (95%) de precisió en les deteccions.
Per a **Bins**, es detecten el 100% de *hard cuts* (deteccions perdudes = 0), notablement major que per a Statistical, mentre que la precisió és més baixa (89%) que la obtinguda per a Statistical.

3.3.6. Notícies i publicitat (*VARIS_06*)

Els programes de notícies solen presentar efectes especials en la transició entre notícies, la qual cosa pot provocar falses deteccions de *shots*, fent minvar la precisió.

La **Fig. 3.7** mostra el resultat de la detecció de *hard cuts* en un programa informatiu, el qual té anuncis intercalats:

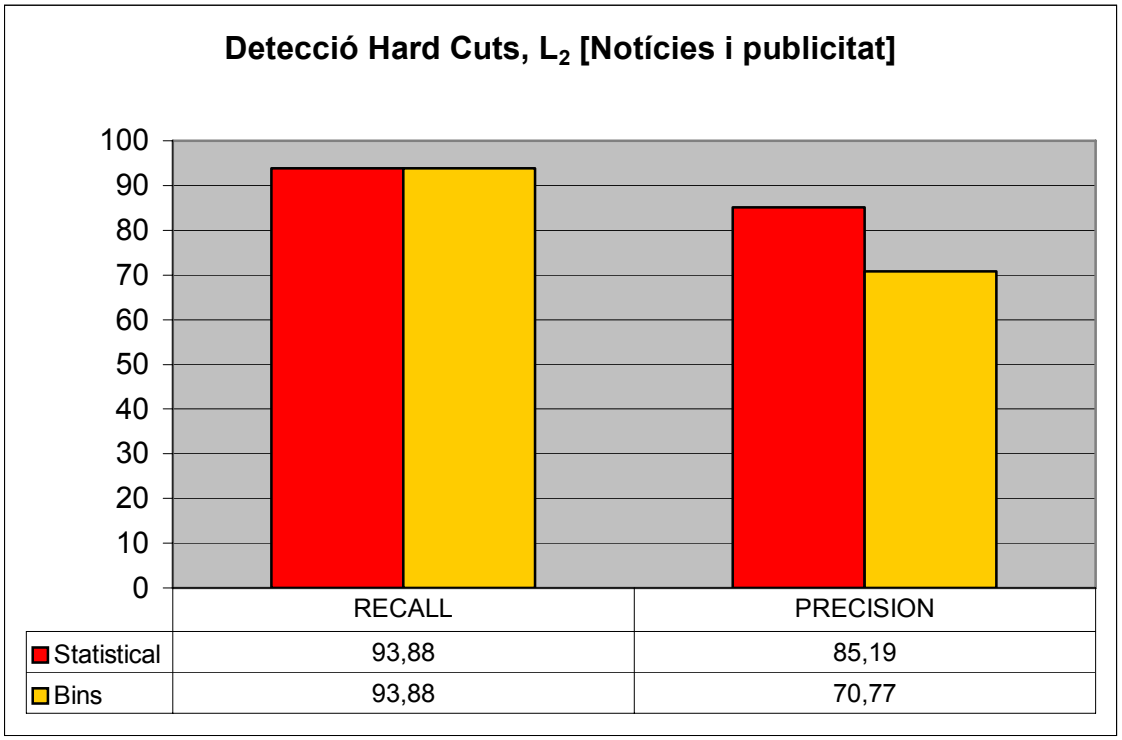


Fig. 3.7 Resultats de la detecció de *hard cuts* en vídeos pertanyents al gènere de notícies i publicitat (*VARIS_06*)

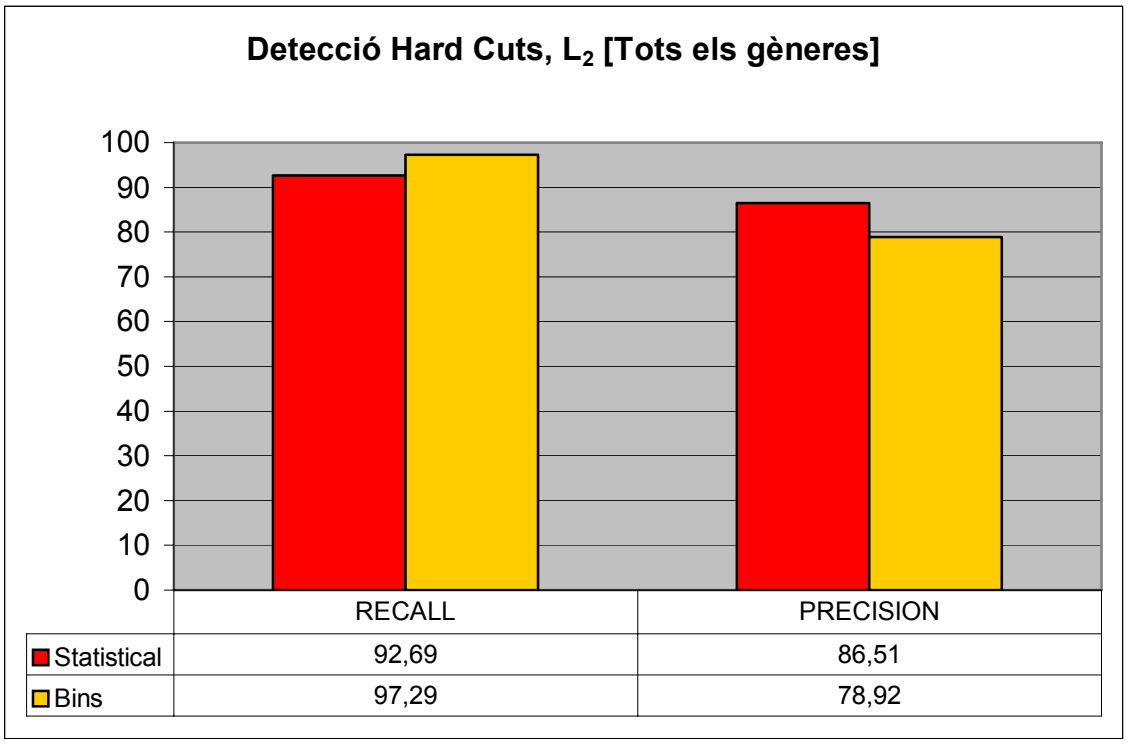
Per a **Statistical**, es detecten un alt nombre de *hard cuts*, això queda reflectit a Recall (94%) i s'obté un bon percentatge (85%) de precisió en les deteccions.
Per a **Bins**, els resultats no són tant bons, es detecten el 94% de *hard cuts*, mínimament menor que per a Statistical, mentre que la precisió també és més baixa (71%) que la obtinguda per a Statistical. Això és degut a l'efecte de les transicions graduals entre *shots*, que provoquen que es detectin incorrectament un gran nombre de *shots*

3.3.7. Tots els gèneres

Finalment, es realitza una valoració global dels resultats obtinguts per a cadascun dels gèneres de vídeo per a decidir quina de les dues característiques de color (Statistical o Bins) és més eficient a l'hora de detectar *shots*.

Els resultats que es presenten són una mitjana aritmètica dels resultats obtinguts per a cada tipus de vídeo.

La **Fig. 3.8** mostra el resultat de la detecció de *hard cuts* per a tots els tipus de vídeo analitzats:



	RECALL	PRECISION
Statistical	92,69	86,51
Bins	97,29	78,92

Fig. 3.8 Resultats de la detecció de *hard cuts* per a tots els gèneres de vídeo

Com era d'esperar, es pot observar el que s'ha vist per a cada tipus de vídeo per separat. Per a **Statistical**, la precisió és globalment és més alta (87%) que per a **Bins** (79%). En canvi el Recall és més gran per a **Bins** (97%), en front d'un 93% per a **Statistical**.

$$\mathbf{RECALL}_{\mathbf{BINS}} > \mathbf{RECALL}_{\mu,\sigma}$$
$$\mathbf{PRECISION}_{\mathbf{BINS}} < \mathbf{PRECISION}_{\mu,\sigma}$$

En base als resultats obtinguts, es considera millor la sobredetecció de *shots* (fer més deteccions errònies) que no pas ser més acurat en les deteccions que es fan (no fer tantes deteccions errònies) si després no es detecten tants *shots*.

És a dir, donar més importància a la pròpia detecció en si, encara que la precisió no tingui un percentatge semblant a un mètode més acurat però amb més pèrdua de deteccions.

Això és així perquè és millor extreure el contingut visual més cops del compte o no extreure'l de forma precisa, que, simplement, no tenir tanta probabilitat d'extreure'l, encara que si s'extreu és més probable que sigui correcte, ja que és més acurat.

Per tant, s'arriba a la conclusió que el mètode de detecció de *hard cuts* més eficient és a partir dels bins de l'histograma HSV extrets amb el ***Scalable Color Descriptor*.**
Així i tot, els resultats obtinguts a partir de μ i σ del canal HSV dels *frames* de les seqüències de vídeo, són també força eficients.

CAPÍTOL 4. CONCLUSIONS I LÍNIES FUTURES

Conclusions

S'han implementat dos descriptors de color, el *Scalable Color Descriptor* (SCD) i el *Group of Frames/Pictures Descriptor* (GoF), definits en l'estàndard de descripció de contingut multimèdia, MPEG-7.
El SCD extreu els bins de l'histograma de l'espai de color HSV i després els codifica mitjançant la Transformada *Wavelet Haar*.

Una vegada extretes les característiques de color, s'han calculat mesures de distància L_2 i χ^2 entre *frames* consecutius que proporcionen la informació necessària per a, aplicant algorismes basats en llindars temporals adaptatius, detectar els *shots* d'una seqüència de vídeo.

S'han presentat un conjunt de resultats per a tots els gèneres de vídeo inclosos en la base de dades segmentada manualment.
En primer lloc, s'ha descartat la mesura de distància χ^2 per a l'avaluació dels resultats, ja que presentava pitjors resultats que la mesura de distància L_2.
Seguidament, s'han avaluat els resultats obtinguts:

- **Statistical**: A partir de la mesura de distància L_2 entre *frames* consecutius per als paràmetres estadístics μ i σ del canal HSV.
- **Bins**: A partir de la mesura de distància L_2 entre *frames* consecutius per als bins de l'histograma extret pel SCD.

L'avaluació s'ha realitzar mitjançant *Recall*, que mesura el percentatge de *shots* no detectats (N_m) respecte el total de *shots* (N_t) i *Precision*, que mesura el percentatge que mostra com d'acurat és l'algorisme a l'hora de no detectar *shots* incorrectes (N_f).

Per a la valoració global del gènere de vídeo s'han extret els següents resultats per a la detecció de *hard cuts* en seqüències de vídeo:

$$\textbf{RECALL}_{\textbf{BINS}}\ \textbf{(97,29\%)} > \textbf{RECALL}_{\mu,\sigma}\ \textbf{(92,69\%)}$$
$$\textbf{PRECISION}_{\textbf{BINS}}\ \textbf{(78,92\%)} < \textbf{PRECISION}_{\mu,\sigma}\ \textbf{(86,51\%)}$$

S'ha donat més pes a l'obtenció d'un *Recall* més alt, enlloc de tenir una precisió més bona.
La qual cosa significa que és millor sobredetectar (fer deteccions errònies) els *shots* d'una seqüència de vídeo, que no pas tenir una precisió millor (no fer tantes deteccions errònies) si després es perden més deteccions de *shots* (no detectar tants de *shots*).

Per tant, el mètode de detecció de *hard cuts* més eficient és a partir dels bins de l'histograma HSV extrets amb el ***Scalable Color Descriptor*****.** Aquest

mètode presenta un *Recall* més elevat que l'altre mètode, és a dir, detecta més *shots*, però és menys acurat, la qual cosa significa que el percentatge de deteccions correctes vers les totals és més baix. Així i tot, com s'ha dit al paràgraf anterior, és millor que es facin sobredeteccions que no que es deixin més *shots* sense detectar.

Tot i així, els resultats obtinguts a partir de μ i σ del canal HSV dels *frames* de les seqüències de vídeo, són també força eficients.

Per altra banda, cal dir que tota la informació que s'extreu en el procés de descripció del contingut de color d'una seqüència de vídeo i la detecció de *shots* s'emmagatzema en fitxers de text (veure Annex).

Línies Futures

Després de la implementació d'un algorisme per a la detecció de *hard cuts*, seria útil, ajustar els paràmetres de l'algorisme de detecció de ***soft cuts*** adequadament per a assolir uns resultats eficients.

Seria útil implementar altres descriptors visuals (color, textura, forma o moviment) definits en l'estàndard MPEG-7, per a incloure'ls al sistema de detecció realitzat en aquest treball.

Per altra banda, la de detecció de *shots* de seqüències de vídeo amb **descriptors d'àudio** és una altra de les possibles tasques futures a desenvolupar. Posteriorment, es podria realitzar una tasca de fusió de l'anàlisi d'àudio amb els algorismes de detecció i descriptors de color implementats en aquest treball. Així, s'obtindria una segmentació dels *shots* a partir de l'anàlisi multimodal -format per l'anàlisi d'àudio i de vídeo- de seqüències de vídeo.

Una altra via que es pot desenvolupar a partir de la tasca realitzada, és a partir de les representacions extretes amb el descriptor GoF de cada *shot* detectat, dissenyar un sistema que sigui capaç de **classificar** el seu contingut.

Estudi d'ambientalització

Aquest és un projecte de desenvolupament d'eines software, i com a tal no té repercussions mediambientals directes.
Com que és obligatori tenir en consideració el medi ambient en la realització de qualsevol projecte, s'analitza el possible impacte indirecte que pugui tenir el projecte com seria l'energia consumida per un ordinador portàtil (de baix consum) i l'ús de consumibles òptics, concretament de DVDs, per a emmagatzemar la base de dades audiovisual segmentada.

REFERÈNCIES BIBLIOGRÀFIQUES

[1] Wang, Y., Liu, Z., and Huang, J.C., "Multimedia Content Analysis Using both Audio and Video Clues", *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12-36, New York, 2000.

[2] Manjunath, B.S., Salembier, P. and Sikora, T., *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley (2002).

[3] Cieplinski, L., "MPEG-7 Color Descriptors and Their Applications", *CAIP* 2001, LNCS 2124, pp.11-20, 2001.

[4] Manjunath, B.S., Ohm, J.R., Vasudevan, V. , and Yamada, A., "Color and Texture Descriptors", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703-715, 2001.

[5] Ferman, M., Tekalp, M. And Mehrotra, R., "Robust Color Histogram Descriptors for Video Segment Retrieval and Identification", *IEEE Transactions on Image Processing*, vol. 11, no. 5, pp. 497-508, 2002.

[6] Dugad, R., Ratakonda, K and Ahuja, N. "Robust video shot change detection", *IEEE Workshop on Multimedia Signal Processing*, 1998.

[7] Zhang, H.J., Smoliar, S.W. and Kankanhalli, A., "Automatic Partitioning of Full-Motion Video". *Multimedia Systems*, vol. 1, no. 1, pp.10-28, 1993.

[8] Zhang, H.J., Smoliar, S.W. and Furht, B., "Content-Based Video Indexing and Retrieval", Cap. 12 en *Video and Image Processing in Multimedia Systems*, Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers (1995).

[9] Ceballos, F.J., *Enciclopedia del lenguaje C++,* RA-MA (2003).

[10] Chapman, D., *Teach Yourself Visual C++ 6 in 21 Days*, Sams Publishing (1998).

[11] Intel Integrated Performance Primitives (IPP), "http://www.intel.com/cd/software/products/asmo-na/eng/perflib/ipp/index.htm".

[12] Intel Open Source Computer Vision Library (OpenCV), "http://sourceforge.net/projects/opencvlibrary/", "http://www.intel.com/technology/computing/opencv/index.htm".

[13] Intel Image Processing Library (IPL), "http://developer.intel.com/software/ products/perflib/ipl/".

ANNEX

Implementació de l'aplicació

Després de l'estudi dels descriptors visuals i de la segmentació manual de material audiovisual, és necessari implementar una aplicació que, amb aquests algorismes, sigui capaç de detectar els *shots* d'un vídeo.
En aquest annex s'expliquen les eines utilitzades per a implementar-ho i l'entorn de programació en el qual s'ha desenvolupat. Així com una explicació de les parts i el funcionament de l'aplicació.

A.1. Entorn de programació

L'aplicació s'ha desenvolupat amb Visual C++ en un entorn Windows.
C++ és un llenguatge de programació d'alt nivell que permet realitzar aplicacions de vídeo en temps real.
És un llenguatge que suporta diferents mètodes de programació, com són la programació estructurada i la orientada a objectes [9].

A.1.1. Programació Orientada a Objectes

Per a implementar l'aplicació en aquest entorn s'ha realitzat utilitzant la Programació Orientada a Objectes (POO).
La POO consisteix en construir components independents d'estructures de dades i rutines les quals estan definides en una classe (d'objectes), de forma que el programa final acaba utilitzant aquests components, que al mateix temps poden interaccionar uns amb altres, de manera paral·lela o jeràrquica.

Els objectes són instàncies d'una classe, que s'utilitza com una variable en un programa, és a dir, la creació d'una instància d'una classe es correspon amb la declaració d'una variable en la programació estructurada, però referint-se a objectes. Un objecte respon a funcions o procediments, que són el principal mitjà de comunicació.

La funcionalitat i la metodologia interna es fonamenta en:

- Encapsulació: En una classe es declaren els tipus de dades i el mitjà de manipular-los (funcions i procediments).
- Herència: Suposa crear classes derivades d'altres ja existents, que hereten els seus tipus de dades i funcions i poden tenir-ne altres de nous. Quan una nova classe hereta propietats de més d'una classe antecessora, s'anomena herència múltiple.
- Polimorfisme: Facilita la programació de funcions i procediments que executaran accions que dependran dels objectes sobre els quals s'apliquin; per exemple, augmentar la mida d'un objecte, independentment de la seva forma.

La **Fig. A.1** il·lustra els avantatges de cada mètode de programació en funció de la complexitat del programa a implementar:

- Disseny no estructurat: Programar sense cap tipus d'estructuració no té sentit. Quan la complexitat del programa és molt petita, pot ser la forma més ràpida de desenvolupar-ho, però serà molt complicat fer modificacions futures. Aquest mètode es pot descartar.
- Disseny estructurat: A diferència del disseny anterior, amb el disseny descendent estructurat és més difícil desenvolupar un programa molt senzill, però si la complexitat augmenta, serà més fàcil desenvolupar-ho.
- Disseny orientat a objectes: És necessari un esforç considerable per a implementar programes petits, però si les aplicacions són grans i hi treballen diferents programadors és la opció indiscutible ja que és més senzill modificar-ho en un futur.

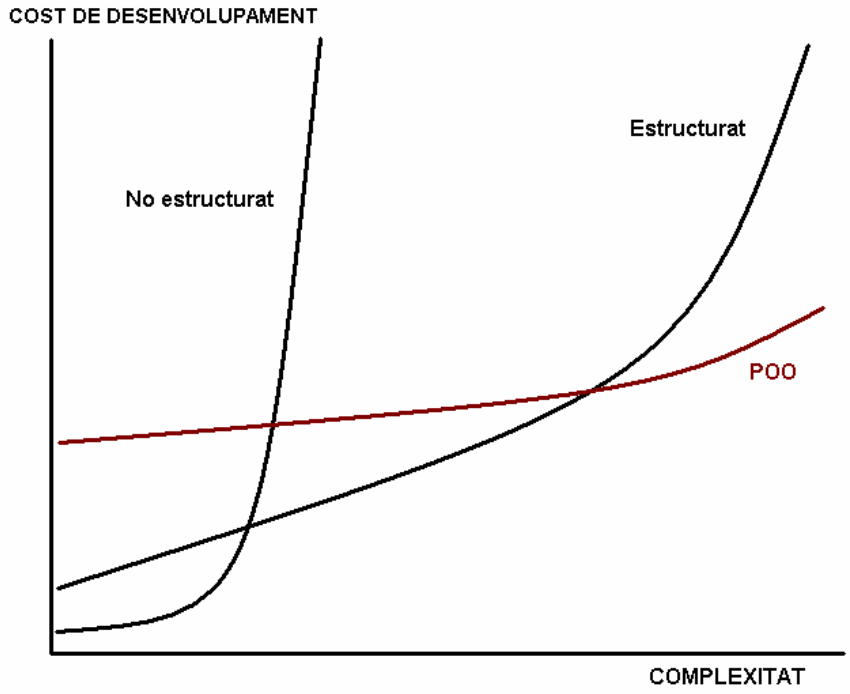


Fig. A.1 Gràfic de les diferents tècniques de programació en funció de la complexitat del programa

A.1.2. Visual C++

La implementació de l'aplicació s'ha realitzat amb l'entorn de desenvolupament *Microsoft Development Environment 2003 (Microsoft Visual Studio .NET)*, concretament amb l'entorn *Microsoft Visual C++ .NET*.

L'entorn permet crear i dissenyar *Graphic User Interfaces* (GUI) amb molta facilitat. També inclou les llibreries *Microsoft Foundation Classes* (MFC) i *Application Program Interface* (API) les quals permeten la creació de

components típics de Windows, com per exemple, diàlegs, botons, barres de desplaçament, gràfics, etc [10] .
A l'entorn s'hi han adherit diferents llibreries útils per al tractament digital de senyals, i concretament de senyals audiovisuals. Aquestes eines es detallen al subapartat següent (A.1.3).

A.1.3. Eines de desenvolupament addicionals

Les eines relacionades amb el processat d'imatge que s'han utilitzat per a implementar els descriptors i per a fer l'aplicació han estat:

- *Intel Integrated Performance Primitives (IPP)*

La d'*Intel IPP* [11] opera en diferents tipus de senyals. Per un part, existeix un paquet de la llibreria que treballa amb senyals unidimensionals (*Volume 1: Signal Processing*) i per altra banda, hi ha un altre paquet de la llibreria que treballa amb senyals bidimensionals usades en el processament d'imatge i vídeo (*Volume 2: Image and Video Processing*).
El rendiment en l'ús de les funcions incloses en cadascun dels paquets, es veu millorat si es disposa d'un processador *Intel Pentium* amb tecnologia MMX. La llibreria està enfocada a les instruccions SIMD (*Single-Instruction, Multiple-Data*), les quals també milloren el rendiment del processament.
Conté funcions que realitzen filtratge, transformació (FFT, DCT, geomètriques), càlcul d'histogrames, comparació amb llindars, operacions aritmètiques i morfològiques, càlcul de paràmetres estadístics, etc.

- *Intel Open Source Computer Vision Library (OpenCV)*

La llibreria de visió artificial de codi obert d'*Intel* (*Intel OpenCV*) [12] conté una col·lecció de funcions en C i classes en C++ que implementen algorismes de processament d'imatge i visió artificial (branca de la intel·ligència artificial interessada en el processament d'imatges del món real).
OpenCV està basat en una versió anterior de la llibreria *IPP*, la *Intel Image Processing Library* (IPL) [13]. Ambdues són compatibles entre elles.
Treballa de forma conjunta amb la llibreria *Intel IPP* proporcionant així un ventall més ampli de possibilitats per al processament d'imatge.

Les funcions que incorpora estan agrupades en diferents branques, com són el processament d'imatge, l'anàlisi de moviment, el reconeixement d'objectes, el calibratge de càmeres, la reconstrucció tridimensional, la *GUI* i l'adquisició de vídeo

A.2. Implementació del software de l'aplicació

A.2.1. Classes implementades

Una classe és un tipus definit per l'usuari que descriu els atributs i els mètodes (funcions i procediments membre) dels objectes que es crearan a partir d'aquesta. Els atributs defineixen l'estat d'un determinat objecte i els mètodes són les operacions que defineixen el seu comportament. Formen part d'aquests mètodes els *constructors*, que permeten iniciar un objecte, i els *destructors*, que permeten destruir-lo. Els atributs i els mètodes es denominen en general *membres* d'una classe.

En la implementació de l'aplicació s'han definit una sèrie de classes:

- *ColorDescriptors*: Proporciona funcions que implementen els descriptors de color.
- *FeatureExtraction*: Proporciona funcions per extreure les característiques dels descriptors de color implementats a la classe ColorDescriptors; també proporciona les funcions que realitzen els algorismes per a la detecció de shots.
- *PlotFeatures*: Proporciona funcions per a representar en una gràfica bidimensional (obtinguda a partir d'un control *ActiveX*) les característiques desitjades per l'usuari, les quals poden ser: la mitjana i la desviació estàndard RGB i HSV, així com l'histograma de color HSV i els histogrames *Average, Median* i *Intersection* corresponents al *Group of Frames Descriptor*.
- *ProcessAVI*: Incorpora funcions per obrir fitxers de vídeo en format AVI, i emmagatzema informació referent al senyal de vídeo carregat.

La **Fig. A.2.** mostra un diagrama de blocs de la interconnexió de les classes implementades:

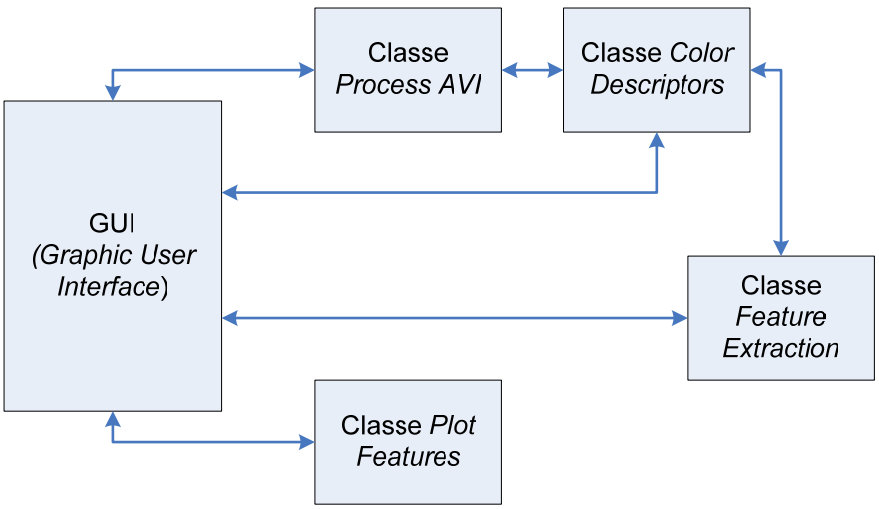


Fig. A.2. Diagrama de blocs de la interconnexió de les classes

A.2.2. Funcionament de l'aplicació

Quan s'obre l'aplicació, es pot veure el diàleg de la **Fig. A.3**, el qual està dividit en cinc parts:

- *AVI*: Carrega i reprodueix fitxers de vídeo en format AVI.
- *Color*: Extreu característiques de color. Descriptors SCD i GoF.
- *Shot Detection Algortihm*: Detecta els *shots* del fitxer AVI.
- *Fast Processing*: Processament directe del fitxer AVI.
- *Graphic of Features*: Representa gràficament diferents característiques estadístiques i diferents histogrames de color.

A continuació es detallaran cadascuna d'aquestes parts.

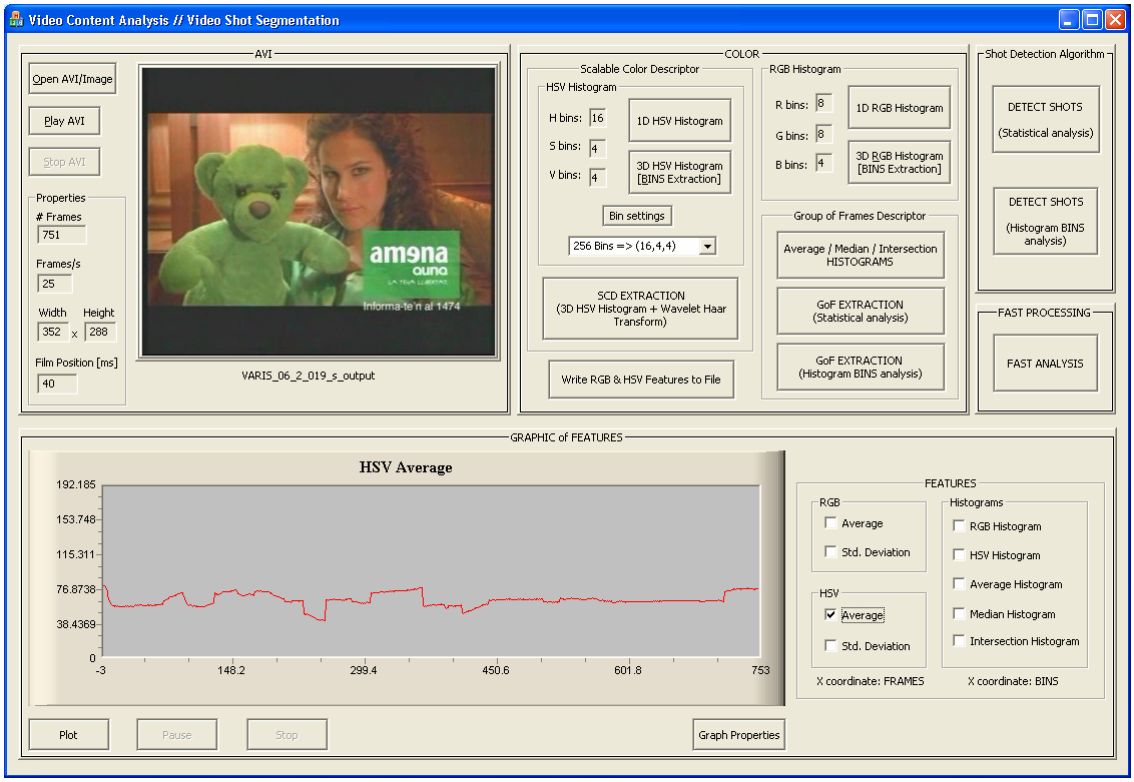


Fig. A.3 Vista general de l'aplicació per a la detecció de *shots* de vídeo

Part *AVI*

El requadre que està situat a la part superior esquerre del diàleg s'encarrega de carregar, reproduir i parar un fitxer de vídeo en format AVI mitjançant els botons que es detallen a la **Fig. A.4**, on també s'observa informació del fitxer que s'ha carregat, com són:

- El número de *frames* del vídeo.
- La taxa de *frames* per segon.
- El tamany de cada *frame* (*Width x Height*).
- La posició actual del vídeo en mil·lisegons.

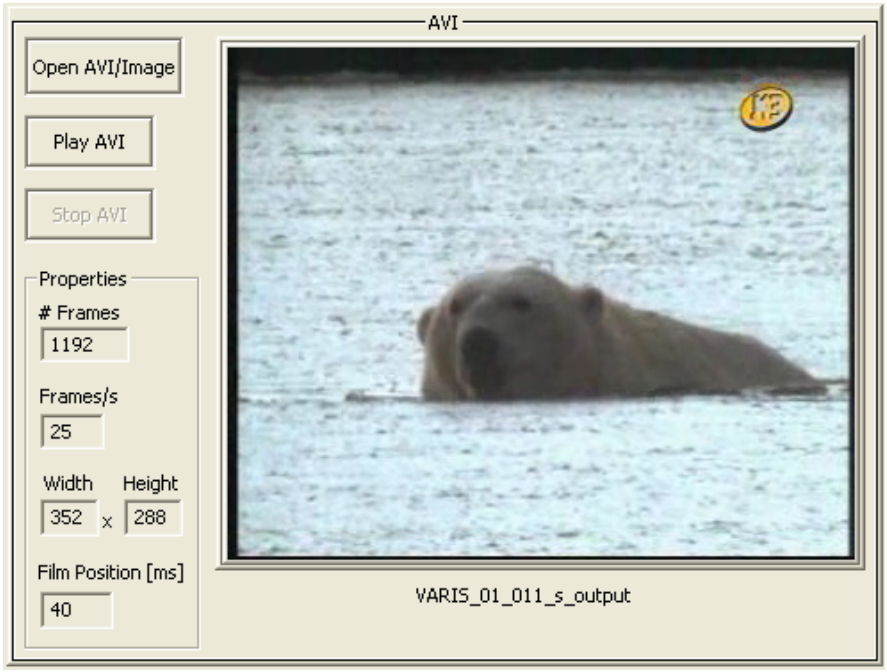


Fig. A.4 Detall de l'aplicació on s'observa el requadre de reproducció i d'informació del fitxer de vídeo carregat

Parts *Color*, *Shot Detection* i *Fast Processing*

Els requadres situats a la part superior dreta, veure **Fig. A.5**, són els que s'encarreguen de fer tot el processament necessari per a poder detectar els *shots* del vídeo que s'ha carregat.

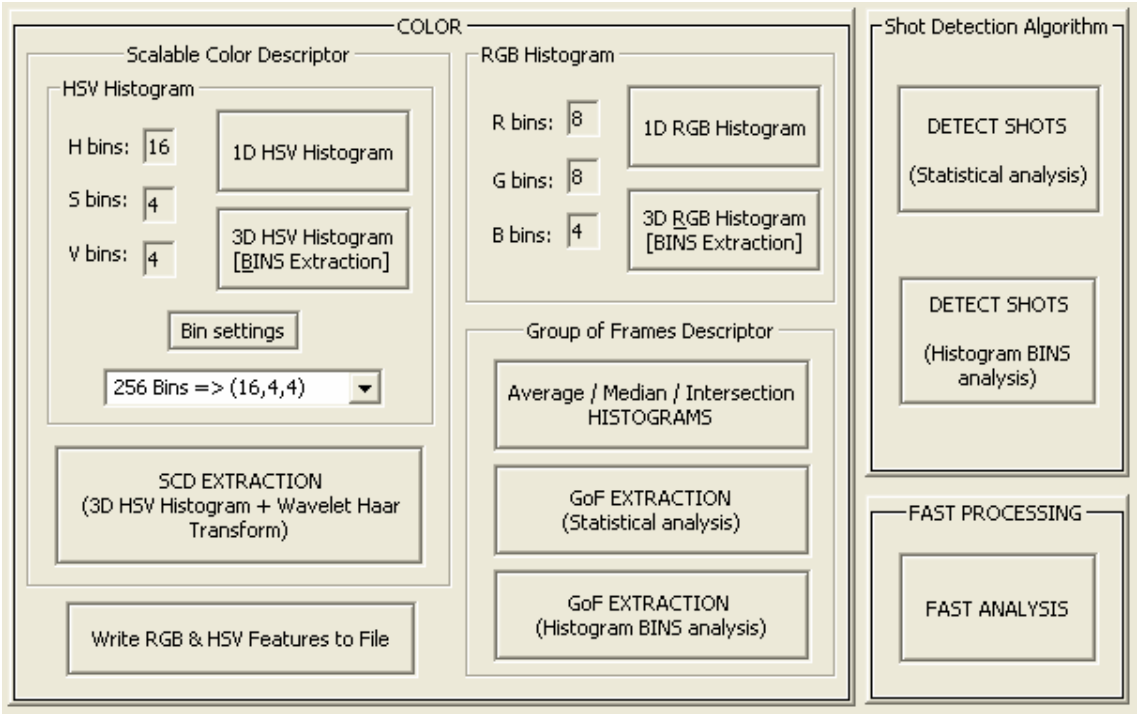


Fig. A.5 Detall de l'aplicació on s'observen els requadres d'extracció de característiques de color, de detecció de *shots* i de processament ràpid

Dins el requadre de *COLOR* es pot calcular l'*Scalable Color Descriptor* per a 16, 32, 64, 128 i 256 bins de l'espai de color HSV definit. En calcular-ho s'extreuen els fitxers amb el següent ordre:

- *VARIS_0X_00X_s_lastframe_3DHSVHist.txt* (1)
  Conté els valors de tots els bins de l'histograma de color HSV de l'últim frame que es processa. També inclou els paràmetres estadístics de la mitjana i la desviació estàndard dels canals H, S i V i una combinació dels tres, HSV.

- *VARIS_0X_00X_s_WaveletHaarTransformsSteps.txt* (2)
  Conté els passos que segueixen els bins de l'histograma de color HSV de l'últim frame per a ser transformats amb la *Wavelet Haar Transform*. La seva utilitat és per a comprovar si s'ha aplicat correctament el procés de transformació.

- *VARIS_0X_00X_s_WaveletHaarTransformResult.txt* (3)
  Conté els valors dels bins resultants d'aplicar la *Wavelet Haar Transform*.

- *VARIS_0X_00X_s_DistanceMeasures_Bins.txt* (4)
  Escriu les mesures de distància L_2, χ^2 i *Histogram Intersection* entre els bins dels *frames* consecutius.

- *VARIS_0X_00X_s_DistanceMeasures_SP.txt* (5)
  Escriu les mesures de distància L_2, L_1, χ^2 i *Histogram Intersection* entre *frames* consecutius de la mitjana i la desviació estàndard (paràmetres estadístics) del canal HSV.

El requadre *RGB Histogram* de *COLOR* permet calcular l'histograma per a 256 bins de l'espai de color RGB definit. En calcular-ho s'extreuen els fitxers amb el següent ordre:

- *VARIS_0X_00X_s_lastframe_3DRGBHist.txt* (6)
  Conté els valors de tots els bins de l'histograma de color RGB de l'últim frame que es processa. També inclou els paràmetres estadístics de la mitjana i la desviació estàndard dels canals R, G i B i una combinació dels tres, RGB.

- *VARIS_0X_00X_s_DistanceMeasuresRGB.txt* (7)
  Escriu les mesures de distància L_2, L_1, χ^2 i *Histogram Intersection* entre *frames* consecutius de la mitjana i la desviació estàndard (paràmetres estadístics) del canal RGB.

Es poden escriure els paràmetres estadístics dels canals RGB i HSV per a cada frame en un mateix fitxer*: VARIS_0X_00X_s_AllFeatures.txt* (8).

Al requadre de *Shot Detection Algorithm* és on s'apliquen els algorismes necessaris per a detectar els *shots* del vídeo. Prement el botó *DETECT SHOTS (Statistical analysis)* es detecten els shots del vídeo carregat utilitzant l'anàlisi

estadístic i s'extreu el fitxer *VARIS_0X_00X_s_ShotsDetected_SP.txt* (9), que escriu els *frames* entre els quals s'ha detectat un *shot*.
El botó *DETECT SHOTS (Histogram BINS analysis)* realitza la mateixa tasca que el botó anterior amb la diferència que l'anàlisi utilitzat per a detectar els *shots* és el dels bins de l'histograma original de cada *frame*. S'extreu el fitxer *VARIS_0X_00X_s_ShotsDetected_Bins.txt* (10), que també escriu els *frames* entre els quals s'ha detectat un *shot*.

El requadre *Group of Frames Descriptor* de *COLOR* permet calcular els histogrames definits al descriptor per a cada *shot* que s'ha detectat a partir de l'anàlisi estadístic i a part de l'anàlisi dels bins de l'histograma original de cada *frame*. En calcular-ho s'extreuen els següents fitxers:

- *VARIS_0X_00X_s_GoF_SP.txt* (11)
  Conté els valors dels bins per als histogrames *Average*, *Median* i *Intersection* de cada *shot* detectat a partir de l'anàlisi estadístic. S'escriuen els bins de cada histograma abans d'aplicar la *Wavelet Haar Transform* i després d'aplicar-la.

- *VARIS_0X_00X_s_GoF_Bins.txt* (12)
  Exactament igual que el fitxer anterior amb la diferència que el càlcul dels histogrames es fa per a cada *shot* detectat a partir de l'anàlisi dels bins de l'histograma original de cada *frame*.

Quan s'ha realitzat el càlcul del *GoF* van sortint a una finestra cada 2 segons els *frames* claus de cada *shot* (*Shot KeyFrame*) que s'ha detectat. A la **Fig. A.6** s'observa un exemple.

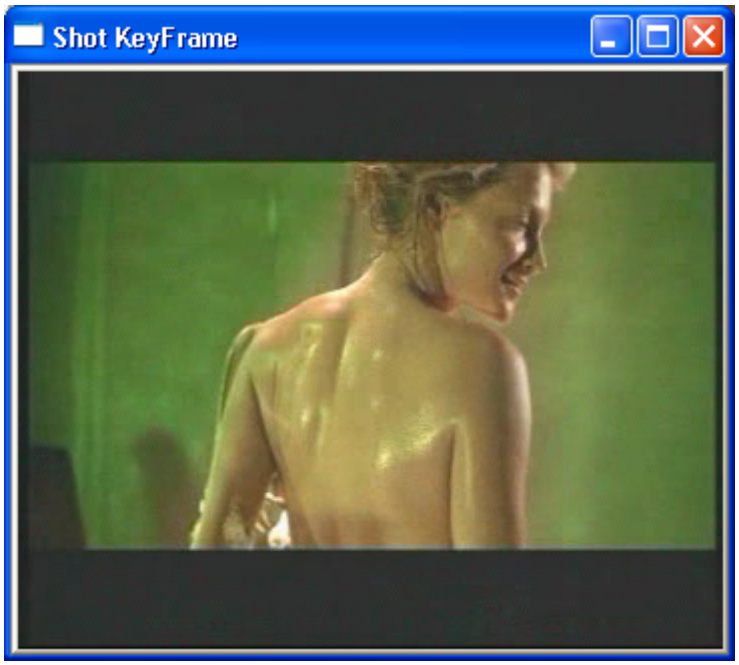


Fig. A.6 Exemple de *Shot KeyFrame* detectat en un anunci publicitari

Finalment, el requadre *FAST PROCESSING* realitza tot l'anàlisi del vídeo carregat. És un mètode ràpid d'extreure els resultats.

Part *Graphic of Features*

La **Fig. A.7** mostra el requadre inferior de l'aplicació, el qual s'encarrega de representar gràficament les característiques que s'han utilitzat per a detectar els *shots*.
Les característiques que es poden seleccionar són:

- Mitjana i desviació estàndard dels canals RGB i HSV.
- Histogrames RGB i HSV de cada *frame*.
- Histogrames *Average, Median* i *Intersection* d'un GoF.

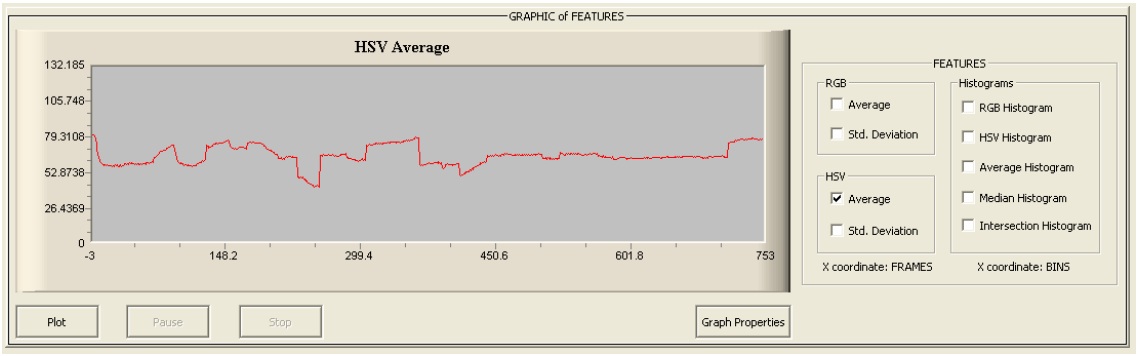


Fig. A.7 Detall de l'aplicació on s'observa el requadre de representació gràfica de característiques

La **Fig. A.8** presenta un diagrama de blocs simplificat de l'ordre que s'ha de seguir per al correcte funcionament de l'aplicació. Els nombres vermells entre parèntesi són els fitxers de text que s'extreuen en cada bloc.

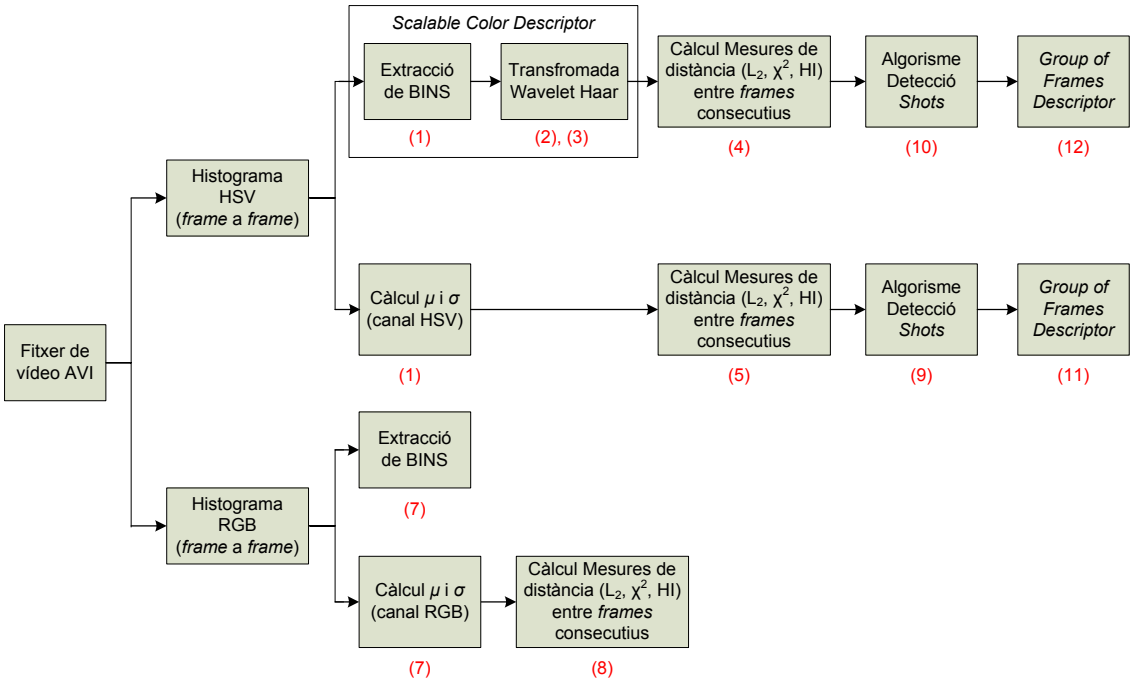


Fig. A.8 Diagrama de blocs simplificat de l'aplicació