Resum del projecte: Classificació i reconeixement de vídeos

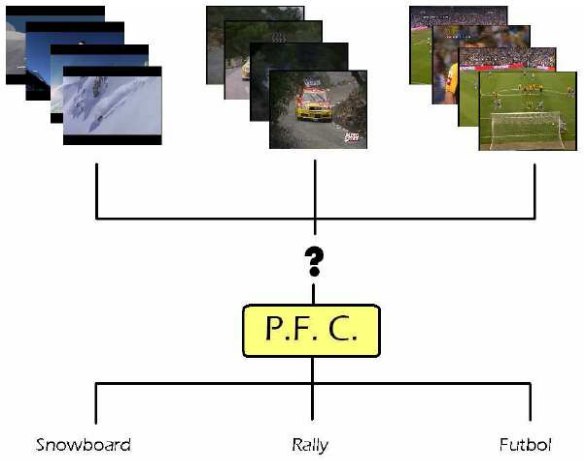
Introducció:

Degut a la proliferació de pàgines dedicades al contigut audiovisual, com per exemple www.youtube.com o www.metacafe.com, i a l'èxit dels programes d'intercanvi de fitxers, com ara l'emule, vàrem creure que seria interessant intentar trobar un mètode que facilités la búsqueda i la classificació dels diferents vídeos a través del seu contingut en lloc de fer-ho segons el seu nom.

Aquesta tasca era una feina bastant complexe degut al gran nombre de classes diferents. Per tal de simplificar-la vàrem decidir reduir les categories de vídeos dins de l'àmbit esportiu.

Objectius:

El principal objectiu d'aquest projecte és aconseguir classificar diferents vídeos d'esports segons la seva categoria.



Per tal de poder assolir aquest objectiu final vàrem decidir emprar un mètode semblant als que utilitzen els cercadors de text: analitzar les paràules visuals que hi aparèixen. Si per identificar un text de cuina un cercador busca paraules semblants a "menjar", "olla", "coure" etc. Per poder identificar un vídeo o una imatge, per exemple, de futbol buscarem diferents paraules visuals com un fragment de la gespa del camp, la pilota o la porteria. Segons la freqüència amb què apareixien les diferents paraules visuals en un vídeo esperàvem poder determinar-ne la categoria.

Descripció dels vídeos:

En aquest projecte vàrem treballar amb 6 classes de vídeos diferents: Fubol, Bàsquet, Rally, Snowboard, BTT i Competicons de Circuit. De cada una d'aquestes

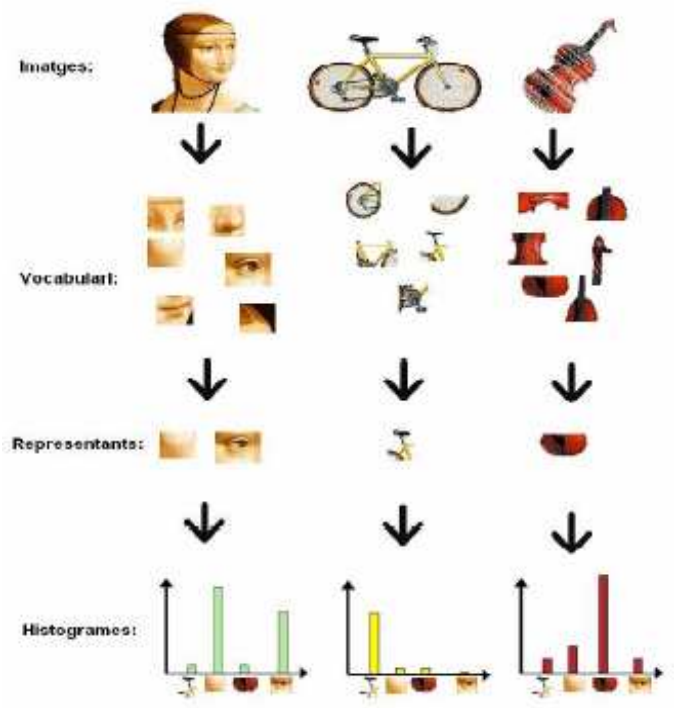
categories vàrem extreuren diferents fragments de vídeo amb una duració total d'entre 8 i 10 minuts per tal de realitzar-hi diferents proves. De cada un d'aquests vídeo vàrem extreuren un frame cada segon per poder-hi treballar posteriorment.

Shots

El primer pas que vàrem realitzar per a poder identificar un vídeo va ser buscar-ne els diferents tipus d'imatge que hi apareixien: els diferents *shots* o escenes. Un *shot* és una seqüència d'imatges consecutives que tenen característiques semblants:

Separant els vídeos en shots vàrem poder eliminar determinades imatges que no aportàven nova informació.

Vocabulari Visual:



Els cercadors de text creen un vocabulari segons el significat de les diferents paraules per tal de poder identificar un document. Per exemple, engloben com a una mateixa paraula les diferents conjugacions d'un verb (canto, cantava, cantarà...) o grups de sinònims (bell, maco, bonic...). En aquest projecte es va fer el mateix però mitjançant paraules visuals: per exemple, es van intentar englobar com a una unica paraula les diferents rodes que apareixien en els cotxes de rally o els diferents fragments de la gespa d'un camp de futbol. A partir de la freqüència amb que apareixien les paraules dels diferents grups dins d'una imatge vàrem crear histogrames de vocabulari que ens permetien tenir una descripció de la imatge.

Detectors de regions

Per identificar les diferents paraules que apareixien en una imatge vàrem decidir adoptar dos mètodes diferents. Vàrem optar per utilitzar detectors de regions (Harris i Mser): programes que obtenen zones característiques d'una imatge; i un Regular Grid: un patró que extreu una regió de la imatge cada cert nombre de píxels.


Imatge Original


Paraules trobaes utilitzant un detector de regions


Paraules trobaes utilitzant un regular grid

Després de realitzar vàries proves i analitzar-ne els resultats vàrem creure que els que millor resultat ens oferien eren el detector MSER i el regular Grid.

Descriptors de regions SIFT

Els descriptors de regions SIFT ofereixen la possibilitat de descriure una regió d'una imatge independentment de la seva posició, angle de visió i lluminositat. Mitjançant aquest sistema obteniem la mateixa descripció de les òptiques d'un cotxe tant si es miraven frontalment com si es feia amb una inclinació de 45º.

Un cop vàrem haver descrit les paraules les vàrem agrupar mitjançant l'algorisme de clusterització "k-means". D'aquesta manera obtindriem els vocabularis que ens servirien per a realitzar els diferents histogrames.

Correspondència entre imatges

Per analitzar si els histogrames que havíem obtingut eren suficientment bons vàrem utilitzar el que es coneix com a "image retrieval". Comparant l'histograma d'una

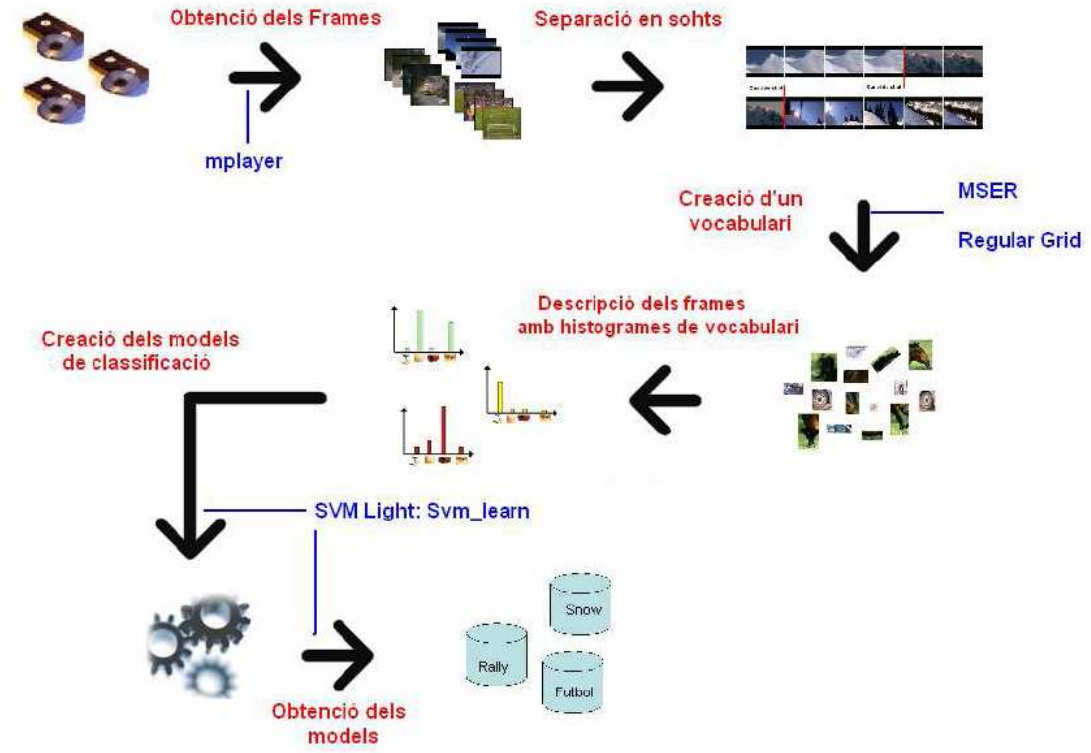
imatge determinada amb la resta d'histogrames, obtenir les k imatges que més s'hi assemblaven.

Mitjançant el mètode anomenat "Precision & Recall" vàrem observar l'eficiència dels dos vocabularis que havíem utilitzat: un creat a partir del detector de Regions MSER, i l'altre creat amb el Regular Grid. Després d'analitzar els resultats vàrem desestimar l'us del vocabulari MSER ja que el seu rendiment era molt inferior al del vocabulari Regular Grid.

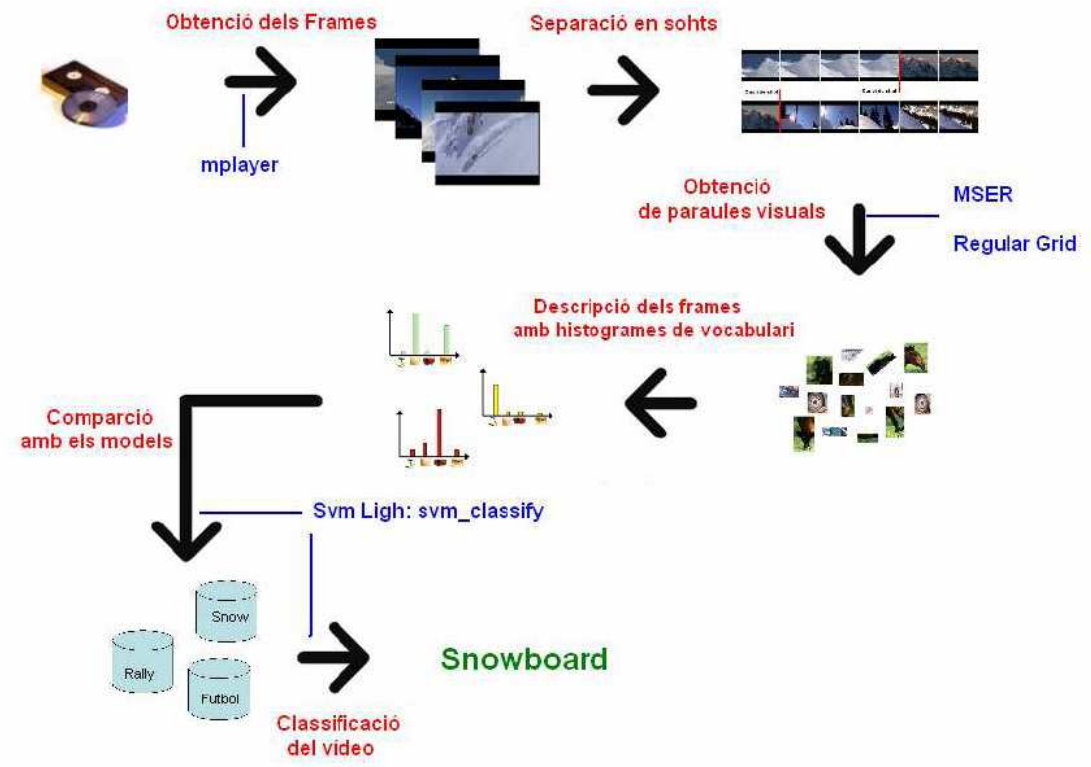
Classificació segons classes

Per classificar un vídeo vàrem decidir utilitzar els histogrames que descrivien els seus frames. Com que cada histograma es podia considerar un vector de valors enters vàrem optar per utilitzar una màquina classificadora de vectors: una Support vector machine o SVM.

Les SVM permeten crear models d'identificació mitjançant un entrenament previ. En aquest entrenament cal introduir-hi un determinat nombre de vectors i indicar-l'hi quins pertànyen a una classe i quins a una altre. Posteriorment la SVM retorna un model que permet classificar els vectors de la categoria entrenada.



A través dels models obtinguts i les SVM es pot classificar un vector histograma segons la seva categoria.



Després de realitzar diferents entrenaments amb varis nombres d'histogrames vàrem aconseguir uns models suficientment bons que ens van permetre classificar correctament 25 de 30 vídeos.

Conclusions

- Els sistemes d'Image Retrieval i el classificador que hem implementat tenen un funcionament lògic ja que els conjunts d'imatges amb els que s'obtenen pitjors resultats són part de classes de vídeo molt semblants i que comparteixen trets en comú com per exemple rally i competicions de circuit
- Aquest sistema de classificació o recuperació de vídeos i imatges no es podria utilitzar en temps real ja que els seus temps d'execució són molt alts. Per exemple, la classificació d'un vídeo amb una durada d'aproximadament mig minut tarda aproximadament uns 50 minuts degut al procés d'extracció i anàlisis dels diferents frames. No obstant aquest procés es podria realitzar offline
- El mateix succeeix amb la creació dels models on el temps que s'ha necessitat ha estat d'aproximadament una setmana.