UNIVERSITAT DE BARCELONA

TREBALL DE FI DE GRAU

Més que mil paraules

Funcionament i estat de la qüestió de la cerca i recuperació d'informació multimèdia basada en el contingut

Autor: Oriol Cantarell Gutiérrez
Tutor: Gema Santos-Hermosa

Agraïments

Hi ha un proverbi africà que diu “Si vols anar de pressa, fes-ho sol però si vols arribar lluny, vés acompanyat”. Per aquest motiu, m’agradaria recordar a algunes persones que m’han acompanyat en el camí fins a l’assoliment d’aquest treball.

Voldria agrair especialment a la meva tutora, la Dra. Gema Santos el temps, l’interès i el suport invertits per a la realització d’aquest treball.

A les meves companyes i amigues, Cristina Campano, Nerea Galván i Agnès Santamarta pel camí que hem compartit durant el grau per poder arribar fins aquí.

I a la meva família, i en especial, als meus pares, ja que sense ells i la seva paciència, suport i ajuda això no hauria estat possible.

A tots ells: gràcies.

Sumari

Sumari d’imatges i figures

Figures

Imatges

Índex d’abreviatures

ASR	Audio Speech Recognition
AVSR	Audio-visual Speech Recognition
CAD	Computer Aided Diagnosis
CBAR	Content Based Audio Retrieval
CBIR	Content Based Image Retrieval
CBMR	Content Based Music Retrieval
IRMA	Image Retrieval of Medical Applications
MIR	Music Information Retrieval
MMIR	Multimedia Information Retrieval
MoCA	Movie Content Analysis
OCR	Optical Character Recognition
QBE	Query By Example
QBH	Query By Humming
QBS	Query By Sketch
SVM	Support Vector Machine

1 – Presentació, objectius i metodologia del treball

1.1 – Justificació de la tria

El creixement exponencial de la documentació multimèdia des de la popularització d'Internet, i en especial, des de l'aparició dels telèfons intel·ligents, ha provocat que la recuperació d'aquests continguts amb metadades associades necessiti metodologies de cerca addicionals i sistemes de recuperació d'informació adaptats a les necessitats d'aquests continguts.

La cerca i recuperació d'informació basada en el seu contingut és un camp d'investigació de plena vigència, polifacètic i que implica un coneixement pluridisciplinari per a la seva aplicació. És un camp que implica conèixer com es descriu, com se cerca i com es recupera la informació, els processos de reconeixement de patrons per ordinador i el seu funcionament, i en alguns casos, la interacció i integració en sistemes més complexos que van més enllà de la simple consulta-resposta per part de l'usuari.

De fet, l'ús d'aquests sistemes està tan integrat en d'altres que tant les cerques com les recuperacions acaben per ser processos interns d'un programari tancat. Per exemple, en el cas de la conducció assistida, la captació i interpretació d'imatges està completament gestionada per l'ordinador de bord del vehicle. Els processos que intervenen en el reconeixement d'objectes en moviment que fa el vehicle són els mateixos que es farien en una cerca d'imatge amb una altra imatge: s'identifica un cos mòbil i en resposta a la tipologia, el vehicle adapta gradualment la velocitat o en cas d'avançament, la trajectòria, per mantenir distàncies de seguretat.

En altres casos, hi ha una interacció directa entre el sistema de recuperació d'informació basat en el contingut i l'usuari que cerca una resposta concreta, a vegades sense ser del tot conscient dels procediments de cerca que ha utilitzat el sistema. Per exemple, l'usuari captura una cançó amb *Shazam*, que li retorna el títol de la cançó i el grup; però la seva intenció final probablement no és la recuperació d'informació en si mateixa, sinó la possibilitat de poder tornar a escoltar-la més tard, així que l'aplicació dóna l'opció o bé de comprar-la en una botiga digital o inclús afegir-la a un servei de música en *streaming*. En definitiva, la cerca i recuperació d'informació multimèdia forma part de moltes tecnologies no només de present, sinó de futur, i que forma i formarà part de diferents serveis de recuperació d'informació multimèdia.

Aquest treball pretén entendre el funcionament d'aquests sistemes i establir un estat de la qüestió sobre la recuperació i cerca d'informació basada en el contingut. En altres paraules, entendre i conèixer l'aplicació i implicacions de l'ús d'aquells sistemes on, en comptes d'introduir una consulta textual, l'usuari introdueix imatges, sons o inclús petits fragments de vídeo per obtenir-ne d'altres o informacions textuals associades.

1.2 – Objectius

L'objectiu principal d'aquest treball és **elaborar un estat de la qüestió sobre les darreres experiències en cerca i recuperació d'informació no textual (imatge, so i vídeo) basada en el contingut**.

Per a la consecució de l'objectiu principal es plantegen els següents 4 objectius específics:

1. **Exposar i discernir el funcionament bàsic dels sistemes de recuperació d'informació no textual.**

Els diferents algoritmes de recuperació d'informació no textual sovint recorren a un grup de conceptes previs, alguns de certa complexitat, que cal conèixer per entendre quins són els avenços reals que es porten a terme pels investigadors. Aquests sovint impliquen fórmules matemàtiques que en demostren el seu funcionament. L'objectiu d'aquest treball no és entrar a analitzar-les, sinó entendre i comentar les implicacions del seu funcionament.

2. **Contraposar la situació actual dels sistemes de recuperació d'informació no textuals als aspectes detectats a partir de l'anàlisi dels conceptes bàsics.**

Un cop els conceptes pressuposats pels articles han estat explicats, aquests es poden contextualitzar dins de la temàtica i així entendre'ls en tota la seva amplitud, analitzar-los i fer-ne una valoració crítica.

3. **Establir quines són les tendències i aplicacions actuals en la cerca i recuperació d'informació no textual a partir de la revisió i l'anàlisi de literatura especialitzada.**

La investigació sovint genera noves tendències i les millora d'acord amb coneixements previs de la matèria. A partir d'una revisió bibliogràfica, es poden establir tendències cap on van els progressos i quines aplicacions tenen aquestes en la vida quotidiana.

4. **Analitzar i fer una valoració crítica dels avenços i de l'estat actual de la qüestió a partir dels coneixements adquirits en el transcurs del grau i de la investigació.**

Amb l'estat de la qüestió elaborat amb la realització i assoliment dels objectius anteriors, s'aportarà alhora un comentari analític i crític amb els coneixements adquirits en el transcurs del grau i durant el procés d'investigació.

1.3 – Metodologia

Aquest treball proposa una estructura que presenta una panoràmica general per poder entendre el funcionament bàsic dels sistemes de recuperació d'informació no textual. Una vegada establert aquest primer marc de referència, es pretén aprofundir una mica en els diferents aspectes que els conformen, entendre quins són els darrers avenços en la matèria i la seva importància i establir d'aquesta manera l'estat de la qüestió.

Principalment, la metodologia d'aquest treball es basa en la realització d'una anàlisi bibliogràfica de diferents fonts de referència en la temàtica i l'especificació de la situació actual mitjançant articles científics publicats durant el darrer lustre. Per exemplificar i entendre les aplicacions, s'han cercat exemples sense limitació temporal, intentant prioritzar alguns que segueixen en vigor.

Es va fer un primer conjunt de cerques simples a la base de dades *Library and Information Science Abstracts* (LISA) a partir de termes com "*Content Based Image Retrieval*", "*Content Based Audio Retrieval*" i "*Content Based Video Retrieval*" amb un filtre limitant la cerca als darrers 5 anys. Aquesta primera cerca tenia per objectiu trobar articles científics sobre la recuperació basada en el contingut de les diferents tipologies que responguessin a la vigència d'un estat de la qüestió. Amb aquesta cerca i amb altres fonts de bibliografia bàsica, es va preparar una primera estructura d'investigació, contextualitzant els articles amb altres fonts més bàsiques i explicatives.

A partir de l'anàlisi d'aquesta primera estructura, es van detectar una sèrie de conceptes bàsics per als articles però de comprensió més complexa. Entre aquests conceptes es troben els Models Ocults de Markov i els seus termes relacionats, els Models de Mescles Gaussianes i els seus termes relacionats, les "*Suport Vector Machine"*, el "*Movie Content Analisis"* i l'"*Audiovisual Speech Recognition"*. Aquests termes s'han cercat en enciclopèdies en línia (com *Wikipedia*) per entendre'ls, i se n'ampliava la informació en cercadors generalistes (*Google* i *Duck Duck Go*). En alguns casos, també s'ha complementat amb cerques des del CRAI de la UB com en el cas de "*recuperación de información textual"*. Simultàniament, es complementava l'estructura d'investigació amb la cerca en altres fonts d'informació, des del CRAI de la UB fins a altres bases de dades especialitzades no contemplades en la primera cerca com *Library, Information Science and Technology Abstracts* (LISTA), amb termes com "*Multimedia Information Retrieval*", "*Speech recognition*", "*Speaker recognition*", "*Vocal Passport*". Novament, el filtratge va ser de 2014 fins a l'actualitat, i es van endreçar per data els resultats de la cerca, encara que en aquests casos, s'ha tingut més màniga ampla pel que fa als marges temporals de les fonts d'informació, ja que sovint eren cerques per explicar i exemplificar aquests termes, relacionats amb els resultats de la primera estructura d'investigació.

També s'ha fet alguna cerca per citació en casos on realment s'ha considerat important en el context de l'article. Per exemple, l'article d'Avery Wang (Creador de *Shazam*) es va trobar sota l'epígraf "l'algoritme de *Shazam*" en el *pre-print* de Gasser, Rosseto i Schuldt (2019). Per més que sigui una font relativament antiga (2006) pels estàndards establerts per aquest treball, es

tracta d'una font molt important, ja que prové d'un director d'una de les solucions d'èxit en el seu àmbit.

Els exemples basats en fonts de divulgació s'han trobat per serendipitat dels interessos de l'autor mitjançant un lector RSS. Després de l'anàlisi del contingut i dels enllaços d'aquests articles per comprovar la seva veracitat, contrast amb altres fonts i funcionament dels sistemes que es presenten, se'n va valorar la seva pertinència i inclusió.

Finalment, els articles tant científics com de divulgació, s'han gestionat amb el gestor de referències bibliogràfiques Mendeley, on es van etiquetar els articles per la tipologia sobre la qual tractava. Per altra banda, es va crear una carpeta per als articles de divulgació, mantenint d'aquesta manera una separació definida entre les dues tipologies d'articles.

2 – Introducció a la cerca i recuperació d'informació basada en el seu contingut

2.1 – Marc contextual de la recuperació d'informació basada en el contingut

Com indiquen Wu, Xiao i Hong (2018), la cerca i recuperació d'informació funciona principalment de tres maneres: cerca de continguts prèviament classificats, la cerca de continguts a partir d'extractes de text o descripcions textuals i la recuperació basada en el contingut.

Cadascun dels sistemes tenen els seus corresponents avantatges i limitacions. En el cas de la classificació, aquesta cerca es pot fer utilitzant una estructura uniformitzada de categories i subcategories, però a canvi ens trobem davant d'un sistema rígid, que requereix un gran esforç i on hi pot haver una gran subjectivitat en la descripció. Pel que fa a la cerca de documents amb continguts textuals, aquesta es fa per comparació de cadenes de text, que n'estableixen una primera comprensió semàntica i permet l'elaboració d'una indexació. En tots dos casos, ens trobem amb la problemàtica de la càrrega de feina pel que fa a la catalogació i indexació i amb descripcions de continguts que poden ser lingüísticament ambigües, inadequades o subjectives.

Vallez i Pedraza-Jiménez (2007) expliquen i posen exemples a aquesta problemàtica. Moltes cerques es realitzen amb la utilització de llenguatge natural per part de l'usuari i aquest fet té repercussions en la recuperació d'informació. Vallez i Pedraza-Jiménez indiquen que la variació lingüística o sinonímia, que defineixen com la possibilitat d'utilitzar diferents termes per transmetre el mateix concepte, provoca silenci documental. Per altra banda, les ambigüitats lingüístiques o polisèmia, que defineixen com la possibilitat de fer diferents interpretacions d'un mateix concepte, provoquen soroll documental. Així doncs, la cerca i recuperació d'informació textual requereix l'anàlisi i establiment d'un terme normalitzat tant en el procés d'indexació com en el procés de cerca que és incompatible amb l'ús del llenguatge natural sense una adaptació.

La recuperació d'informació basada en el contingut -o *Multimedia Information Retrieval* (MMIR)- gestiona tant la descripció com la recuperació del contingut a partir del mateix contingut de documents de diferent tipologia i respon a necessitats d'informació que difícilment es podrien satisfer amb la cerca textual. Precisament aquest és el principal avantatge de la recuperació d'informació basada en el contingut, que responen als punts febles dels mètodes de cerca anteriors: si no hi ha una cerca textual, no poden haver-hi ambigüitats pel que fa al llenguatge. Per aquest motiu, aquesta recuperació permet una anàlisi objectiva, profunda i exhaustiva del contingut cercat. Un exemple d'aplicació on es pot comprovar la importància de l'eliminació d'ambigüitats és el que presenten Agustí, Valiente i Carretero (2003).

2.2 – Diferències i problemàtiques de la recuperació d'informació multimèdia respecte a la textual

La cerca i recuperació d'informació basada en el seu contingut també té inconvenients: és un sistema de cerca relativament recent i encara relativament imprecís en alguns aspectes, que requereix sistemes de reconeixement per computador especialitzats i econòmicament costosos. A més, cal un document model, una referència a l'hora de cercar, el que implica una idea de cerca molt clara i que probablement s'ha obtingut d'una altra cerca anterior amb algun dels mètodes prèviament explicats. Això se sol denominar a la bibliografia com a Consulta Per Exemple (*Query By Example* o QBE) (entre d'altres, Robles Sánchez, 2004; Ponceleón i Slaney, 2011; Gasser, Rosseto i Schudlt, 2019).

No només és més complicada la cerca, sinó també la seva interpretació i el resultat ofert pels sistemes de recuperació d'informació. Cal tenir en compte que en la presentació dels continguts textuals, elements com la tipografia, tipologia i presentació i organització del text marquen la importància dels continguts. En ser els continguts textuals aquells en els que habitualment es recupera la informació, aquests aspectes formals són més coneguts i assumits tant per qui recupera la informació com pels sistemes de recuperació d'informació. En el cas de la recuperació d'informació basada en el contingut aquests aspectes s'han hagut d'identificar de nou, se'ls ha hagut de donar un pes en el moment de valorar-lo com a resultat i, per tant, s'ha hagut de treballar en la precisió del que s'obtindrà finalment (Wu, Xiao i Hong, 2018).

Però a més, la seva representació un cop recuperada també és més complexa: són documents que computacionalment ocupen més espai i dels quals és més difícil trobar un element concret. I a més, ja tradicionalment la descripció d'un contingut audiovisual implicava la seva visualització completa, requerint un reproductor compatible amb les seves característiques i, per tant, té una difícil representació com a resultat.

Existeixen altres problemàtiques, la més important és el que es coneix com a buit semàntic (Ponceleón i Slaney, 2011 i Bognadov et al., 2013). Aquesta problemàtica consisteix en la dificultat per entendre la informació i el seu significat en conjunt: com més gran és la quantitat i abstracció de les dades que conté un determinat document més difícil és la comprensió dels seus continguts. Per aquest motiu, l'anotació automatitzada segueix sense ser prou precisa especialment en els casos on el buit semàntic és més gran, tot i ser una temàtica molt investigada. I en conseqüència, la recuperació de documents multimèdia segueix depenent, en gran mesura, del seu etiquetatge manual amb metadades o de canviar el procediment de cerca i recuperació amb les diferents alternatives que es presenten al llarg d'aquest treball.

2.3 – Per què les metadades són insuficients per a la descripció de continguts?

La recuperació d'informació basada en el seu contingut és un camp important per a la recerca, ja que la velocitat de generació de documents supera en molt la capacitat per descriure'ls adequadament. Cal tenir en compte que l'aparició d'internet va marcar un punt d'inflexió en la generació de documents digitals, però que ha estat amb l'era dels dispositius mòbils on realment ha augmentat la producció de documents multimèdia convertint, de facto, als seus propietaris en productors. Així i tot, segueix existint la necessitat de filtrar i recuperar aquelles informacions més rellevants en tipologies documentals no tradicionalment associades a la producció per part d'usuaris. En definitiva, calen mètodes més efectius per descriure i gestionar amb garanties tota aquesta documentació sota el risc de perdre-la per sempre, per la seva volatilitat.

Ponceleón i Slaney (2011) assenyalen que per recuperar de manera precisa una determinada escena d'una pel·lícula consultada en text per l'usuari, cal haver-la descrit abans manualment i que, tot i que només cal fer-ho una vegada, és un procés llarg, i sobretot, costós. A més, mentre s'està fent aquesta descripció s'estan fent moltes més pel·lícules fetes per professionals, però també s'estan fent vídeos per gent amb un dispositiu mòbil. El més barat, eficient i objectiu seria que, a partir dels mateixos continguts del document es pogués recuperar, almenys fins que es decidís fer una descripció més exhaustiva.

Per tot plegat, l'ús de metadades automatitzades ajuda a la recuperació de continguts fins que aquests puguin ser descrits més exhaustivament, però són clarament insuficients. En el cas de les metadades textuals, com recullen Ramezani i Yaghmaee (2016) arran d'un article de Zhai (2013), pot ser poc precís, determinat de manera poc evident o amb etiquetes irrellevants. En altres casos, com comenten Bognadov et al. (2013) pel cas de la música, el model de la llarga cua perjudica l'etiquetat manual d'aquelles cançons menys escoltades: les cançons més escoltades estaran més i millor etiquetades que aquelles que no són tan escoltades pel simple fet que són més conegudes. Per aquest motiu, sistemes d'anotació de metadades automatitzats, com el que proposen Pelka, Nensa i Friedich (2018), en aquest cas per a les imatges, busquen no només solucionar aquesta qüestió, sinó establir criteris objectius a través del contingut. Val a dir que, en aquest cas concret, el conjunt d'imatges és limitat, amb un productor clarament definit i que hi ha un llenguatge controlat previst amb anterioritat. Així doncs, sistemes més complexos, requereixen solucions més complexes.

Per tant, l'etiquetat amb metadades de documents no textuals ha de passar de ser un element de descripció bàsic a ser un element descriptiu complementari i un sistema d'anotació i descripció documental automatitzat molt més complet, en els casos on això sigui possible.

3 – Com funciona a nivell bàsic la recuperació d'informació basada en el contingut

La recuperació d'informació basada en el seu contingut depèn de la tipologia del document. Hi ha diferents tipologies de documents multimèdia que van des dels gràfics o sonors, fins als que són combinacions dels anteriors, com el vídeo. Igualment important és recordar el fet que, encara que l'input en la cerca dels continguts sigui no textual, no implica que la sortida hagi de ser-ho per força, amb independència de la seva tipologia.

Per poder treballar amb el contingut de la informació cal traduir-lo de manera estructurada a un llenguatge matemàtic perquè pugui ser processat pels ordinadors. A aquesta traducció se la coneix com a algoritme. En general, els algoritmes que s'apliquen a la recuperació d'informació basada en el seu contingut segueixen procediments similars: sintetitzat del contingut en una fórmula o representació matemàtica, que generalment responen a models probabilístics, i comparació amb elements ja existents als quals se li ha aplicat la mateixa fórmula.

Segons Barrios i Bustos (2011), existeixen dos grans processos en el procés de cerca i recuperació de vídeos basat en el seu contingut: la descripció prèvia dels continguts preexistents i la cerca per similitud.

3.1 – Descripció del contingut

Seguint amb les consideracions fetes per Barrios i Bustos (2011), en primer lloc, cal descriure el contingut dels documents de la base de dades que incorpori el sistema de recuperació basada en el contingut, amb independència de la seva naturalesa i característiques. Aquesta descripció, realitzada a partir de diversos factors, actua com a identificador d'aquell contingut i és un procés que cerca la uniformització d'elements per a la seva posterior cerca per similitud. Aquests descriptors es detallaran posteriorment en els capítols 4.2 – Elements descriptors que hi intervenen per les imatges i 5.2 – Elements descriptors que hi intervenen pels documents sonors.

El procediment de descripció variarà segons la seva tipologia, però és necessari per a poder fer posteriorment les cerques per similitud: si el sistema de recuperació d'informació no sap com són els continguts ja existents, no es poden comparar amb els que es puguin entrar i no pot haver-hi recuperació. Aquesta descripció cal fer-la amb els mateixos algoritmes que posteriorment s'aplicaran en la cerca, per uniformitzar els elements de comparació.

Ras i curt, per obtenir aquests descriptors es transforma el contingut del document multimèdia a un model matemàtic que detecta i definirà els seus descriptors.

3.1.1 – Transformada de Fourier

La transformada de Fourier és una expressió matemàtica complexa clau en les comunicacions modernes, ja que permet convertir en ona –en altres paraules, digitalitzar- qualsevol mena de document (Nieto, 2018; Condliffe, 2016). Aquest fet per tant és molt important si no essencial perquè implica la possibilitat d'uniformitzar qualsevol document digital a un senyal en forma d'ona. Com la resta de transformacions i expressions matemàtiques, és una abstracció i simplificació de la realitat i això comporta limitacions com la pèrdua de fidelitat respecte a un document original.

Aplicat al camp d'estudi d'aquest treball ens trobem, per exemple, que la transformada de Fourier s'aplica tant en el càlcul de la mescla dels colors en la digitalització d'una imatge com en la reducció del soroll de fons per la identificació de cançons. Així doncs, determina numèricament els descriptors corresponents.

3.1.2 – Transformada Wavelet

La transformada Wavelet, *Wavelets* o d'ondetes és una expressió matemàtica similar a l'anterior, amb la diferència que si en el cas de la Transformada de Fourier es tractava el senyal per freqüències, Wavelet ho fa amb major precisió mitjançant aproximacions successives i, per tant, pot retornar informació temporal. S'aplica en la identificació i descripció de patrons i pel seu nivell de detall resulta útil en el reconeixement biomètric.

Segons Agustí, Gonzalez, Miguel et Al. (2014), els *Wavelets* són una representació numèrica que permet establir un criteri de similitud entre un element d'entrada i un d'emmagatzemat en una base de dades. Aquesta representació, segons aquests autors "són un complement bàsic per la caracterització i indexació d'una base de dades". A aquests descriptors se'ls hi dóna diferent importància segons el sistema de recuperació, fet que variarà els resultats que ofereixen.

3.2 – Cerca per similitud

Els sistemes de recuperació d'informació basats en el contingut demanen un document de referència per establir una comparativa entre la mostra i els elements emmagatzemats. Fruit d'aquesta comparació, s'obté un índex o valor de referència –en aquest cas, entès com a índex matemàtic, generalment un valor decimal de 0 a 1-. El document amb l'índex més elevat esdevé el més semblant, i per tant, és el que es retorna com a resultat.

En aquest sentit, la tesi de Robles Sánchez (2004) selecciona 5 fórmules d'un conjunt de 14 proposades prèviament per Haralick et al. (1998) L'efectivitat de les fórmules proposades és on radica la clau i el valor real de l'algoritme, i tot i que la seva utilització s'estudia prèviament a la seva implementació, la seva efectivitat a vegades només s'esbrina amb l'ús.

3.2.1 – Model Ocult de Markov

Els Models Ocults de Markov són grafs que permeten trobar paràmetres ocults a partir d'altres paràmetres coneguts mitjançant la interrelació entre ells. Dins de l'ús de grafs, és el model estadístic amb incògnites més bàsic. S'utilitza tant per al reconeixement de veu, com pel reconeixement òptic de caràcters (OCR), relacionat amb la recuperació d'imatges basades en el seu contingut.

En el cas dels continguts d'àudio, amb els models ocults de Markov, una paraula de dues síl·labes (per exemple, casa) es divideix en trifonemes (/cas/ i /asa/) i per combinatòria, s'extrauria la paraula inicial. D'aquesta manera, es fa una interrelació entre models de llenguatge preestablerts i els models acústics rebuts (Ponceleón i Slaney, 2011). Ràpidament podem veure que, amb la utilització d'aquest model, es pot arribar a una comprensió semàntica de les paraules com a entitats aïllades, però no en conjunt.

3.2.2 – Model de Mescles Gaussianes

Els models de mescles són models probabilístics que permeten la localització de patrons, i a partir d'aquests, la predicció i aprenentatge no manual d'un determinat contingut. La intenció amb aquests patrons és poder-los tractar matemàticament. Cada model té un determinat patró matemàtic. En aquest cas concret, el model de mescles gaussianes és el model de mescles més simple i permet la detecció d'àrees amb característiques comunes, el que s'entén per clústers o agrupacions. (Pedregosa et al., 2011)

Per exemple, en el cas d'una imatge, el color es podria interpretar com un model de mescles gaussianes, però cal simplificar-ho. Per això, en molts casos, s'aplica després d'utilitzar el Model Ocult de Markov que fa una escala de grisos per combinatòria (Bovik, 2005). Així doncs, es fa una primera simplificació i combinació de colors i el Model de Mescles Gaussianes els agrupa.

3.2.3 – Models fusionats

Els models presentats només són els més coneguts, utilitzats i mencionats a la bibliografia, però existeixen més models estadístics. Sovint, aquests models es combinen, fent que la sortida o *output* d'un dels algoritmes alimenti a un de posterior en el transcurs de la seva execució. Aquest fet es coneix com a models fusionats (Ponceleón i Slaney, 2011)[1].

En el cas que ens ocupa, l'ús combinat del Model Ocult de Markov i el Model de Mescles Gaussianes esdevé un model fusionat d'ús força comú en la recuperació basada en el contingut. Una possible explicació a l'ús habitual d'aquesta combinació concreta és que són models que es complementen entre si i són coneguts pels enginyers i, per tant, és més fàcil de preveure com funcionaran quan es combinin.

Per una banda, són models estadístics relativament simples per als enginyers, transversals pel que fa a la seva aplicació en les diferents tipologies documentals i ja bastant coneguts i treballats amb anterioritat. Però per l'altre, potser es troba a faltar una major especificitat o algoritmes d'aplicació més exclusiva a una tipologia, que podrien millorar la precisió del resultat.

Aquests models fusionats tractaran als models entrants i s'inclouran en les màquines de vectors de suport, perquè els sistemes aprenguin els patrons i siguin més precisos.

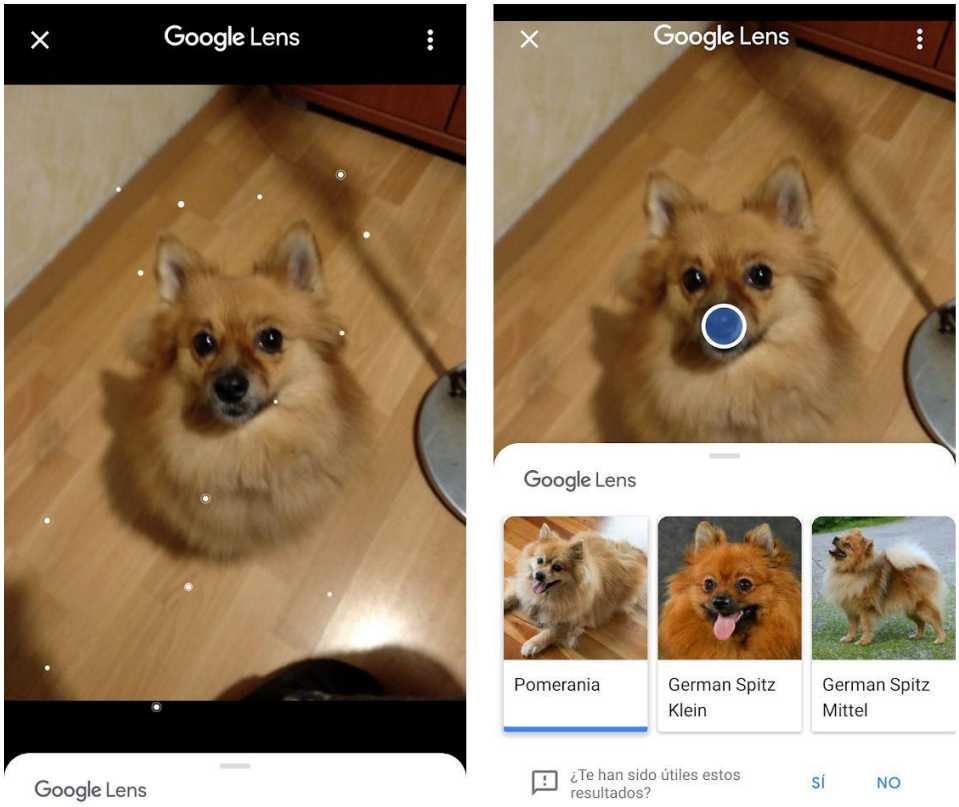
3.2.4 – Màquina de vectors de suport (SVM)

Les màquines de vectors de suport (en anglès, *Suport Vector Machines* (SVM)) són conjunts d'algoritmes, generalment models fusionats, que permeten l'aprenentatge computacional. Amb el conjunt de documents multimèdia, aquests es poden classificar mitjançant categories amb uns intervals mitjans, que es compararan amb les noves mostres amb les quals es pot preveure quina serà la categoria del document entrat. Amb l'entrada de nous valors, aquests valors mitjans varien fent que la màquina aprengui i millori la precisió.

Sovint, molts sistemes de recuperació d'informació basats en el contingut pregunten a l'usuari per la rellevància dels resultats: es pot apreciar un exemple pràctic en la Imatge 1 pel cas de les imatges i també hi ha casos recollits en la bibliografia pel cas del so (Bognadov et al., 2013).

[1] En el cas d'aquests autors, fan referencia a les aplicacions dels models i no als models en si, tot i que amb l'exposició plantejada en els exemples, deixen clar la seva interacció entre si

4 – La recuperació d'imatges basada en el contingut (CBIR)



Imatge 1: Exemple d'ús de CBIR amb Google Lens.
Donada una imatge, identifica l'element predominant de la imatge (Esquerra) i retorna el resultat més semblant destacat amb una franja en blau i altres de similars (Dreta) (Elaboració pròpia)

4.1 – Definició i conceptes bàsics

Tot i que no hi ha una única definició acadèmica o formal, per Recuperació d'Imatges Basada en el Contingut (CBIR per les sigles de *Content-Based Image Retrieval*) s'entén el conjunt de tècniques utilitzades per a l'obtenció d'informació a partir d'un document gràfic donat. En aquest camp de recerca, l'usuari del sistema que l'incorpora proporciona una imatge perquè el sistema li retorni una altra de característiques similars. Això es fa prescindint del significat que pugui tenir la imatge i recuperant característiques com el color (incloent-hi aspectes relacionats com el matís, la lluminositat o el cromatisme), la textura, la forma o les vores. Alguns algoritmes poden aprofundir en altres aspectes com la posició i orientació de l'element destacat en la imatge.

Les imatges van ser un dels primers elements multimèdia del qual se'n van identificar i extreure elements del seu contingut, ja que els patrons de reconeixement d'imatges eren més simples que el de la resta d'elements multimèdia i havien estat anteriorment investigats, per bé que inicialment no van tenir gaire èxit comercial segons Ponceleón i Slaney (2011). A més, com posteriorment indiquen els mateixos autors, el seu buit semàntic era menor i era més fàcil omplir-lo amb informació contextual. Ho exemplifiquen assenyalant que l'any 2001, Google ja

permetia la cerca d'imatges a partir de continguts textuals, i l'associació entre la imatge i el seu contingut es feia aprofitant elements textuals que es trobaven en la mateixa pàgina que la imatge per tancar el buit semàntic i fer la cerca més precisa.

4.2 – Elements descriptors que hi intervenen

Diversos descriptors poden intervenir en l'anàlisi de contingut necessari per a la recuperació de les imatges. De fet, l'ús de diferents descriptors simultanis milloren la precisió del contingut recuperat, però cal donar un pes específic a cadascun d'aquests factors i en complica l'algoritme final. (Wu, Xiao i Hong, 2018)

La correcta anàlisi dels elements descriptors en un entorn estàtic és clau per a l'aplicació posterior en entorns dinàmics (vídeo i altres imatges en moviment). El moviment afecta elements com la posició de l'objecte a la imatge o la percepció de les textures d'un *frame* (que és com s'anomena cadascuna de les imatges estàtiques que conformen el moviment del vídeo) a un altre, el que s'entén com a variació espaciotemporal (Ponceleón i Slaney, 2011).

Aquests elements descriptors poden tenir diverses representacions matemàtiques: des de la representació hexadecimal del color a la representació en matrius de vectors de co-ocurrència en el cas de les textures. En aquest apartat fem un exercici d'abstracció per entendre el seu funcionament sense entrar en aspectes matemàticament massa complexos.

4.2.1 – Color

El color és la interpretació humana de la llum amb diferents longituds d'ona. L'ull humà només pot interpretar un conjunt limitat de colors, que és el que es coneix com a espectre visible. La percepció d'un color pot variar d'una persona a un altre.

La llum prové de la variació en la il·luminació de tres colors bàsics: el Vermell, el Verd i el Blau (o RGB, per les seves inicials en anglès). Aquesta simplificació és el que permet als ordinadors treballar amb els colors, però no és com els humans perceben el color, sinó una interpretació. Així doncs, existeixen interpretacions alternatives al sistema RGB com poden ser el CMYK (Cyan, Magenta, Yellow, Key) o d'altres basats en el matís del color com és el sistema HSV (Hue, Saturation, Value).

4.2.2 – Textures i patrons

Entenem per textura en aquest context la percepció visual de les rugositats i impureses d'un determinat cos. En el cas de la recuperació d'imatges basada en el contingut, se sobreentén que un cos amb textures semblants determina un objecte, i per aquest motiu, ajuda a reconèixer els objectes que apareixen en una determinada imatge.

En general, per detectar les textures es fa una transformació de la imatge en una escala de grisos amb els models que hem descrit anteriorment. El resultat és el que es coneix per histograma.

Amb la detecció dels patrons estadísticament més repetits, mitjançant l'ús de Models Ocults de Markov, es creen un conjunt de vectors que acaben conformant una matriu de co-ocurrència, que s'utilitzarà en la detecció de similituds amb les textures i patrons dels elements ja existents.

4.2.3 – Vores i forma

Amb la detecció de les vores, i en especial, dels vèrtexs dels objectes que apareixen en una imatge es pot determinar que està en primer i en segon pla i la seva forma i àrea dins de la imatge (Wu, Xiao i Hong, 2018). En cas d'haver-hi més d'un objecte en una imatge es pot interpretar que són i quin és el principal a partir de les dimensions de la seva àrea.

Posteriorment, la comparació de vores servirà per a l'elaboració de vectors de moviment en les comparatives entre dues imatges consecutives pròpies del treball amb documents audiovisuals, que resulten especialment interessants en camps com el de la conducció autònoma.

4.2.4 – Posició-regió dins de la imatge

Wu, Xiao i Hong (2018) proposaven, a més de la detecció de les vores, donar un "pes" específic a determinats píxels segons la seva posició dins de la imatge. Amb aquestes delimitacions podien determinar no només la forma (que hem explicat anteriorment) sinó que també se li pot donar un pes i un valor per determinar si és l'objecte principal de la imatge o la seva importància relativa en cas de no ser-ho.

Imatge 2- Exemple d'ús de CBIR amb Pinterest
En aquest cas, es pot comprovar com hi ha similituds pel que fa al color i la textura de les propostes similars. En aquest cas, també es pot cercar per determinades àrees de la imatge i es proposen enllaços i informacions contextuals. Font:<https://pin.it/kkaxjmlqs7auap> [Consulta 29 maig 2019]

4.3 – Aplicacions

Robles Sánchez (2004) en la seva tesi doctoral fa una primera proposició de classificació de les aplicacions de la recuperació d'imatges basada en el contingut, dividint-les en set categories: entreteniment, preparació de documents, aprenentatge, gràfics i publicitat, cerca de logotips aplicacions en medicina i seguretat. Aquesta classificació segueix sent vigent amb matisos, ja que potser els noms no són els més representatius en alguns casos i en d'altres, com en l'entreteniment, encara no hi ha aplicacions prou evidents.

Per aquest motiu, es proposa una petita classificació basada en la de Robles Sánchez on relacionarem algunes de les aplicacions actuals més populars i articles de la bibliografia en els diferents apartats proposats.

4.3.1 – Preparació de documents

Un exemple de preparació de documents és el que expliquen Moreno-Schneider, Martínez i Martínez Fernández (2016) pel cas de l'edició d'una peça informativa. Expliquen que una vegada editada la notícia, aquesta s'ha de poder recuperar, fins i tot amb la informació associada que pugui contenir per poder crear noves peces informatives i que la cerca posterior "ha de ser simple, ràpida i transparent". Quan parlen de "simple", indiquen que no es poden combinar tipologies de cerca, és a dir, fer una cerca textual i alhora una altra basada en el contingut, en especial si la cerca ha de ser ràpida, per més que el resultat si pot combinar-les.

4.3.2 – Aprenentatge

Els continguts desconeguts poden ser utilitzats per a la comparació amb altres imatges per a la seva posterior identificació i descripció textuals. És el cas de la cerca d'imatges de Google o del servei de *Google Lens*.

De fet, en l'exemple de la Imatge 1, veiem un exemple d'ús de Google Lens en el que el procés d'identificació que s'utilitza permet resoldre la pregunta de "que estic veient concretament?". Aquest element permet el descobriment que, no només apareix un gos, sinó quina raça de gos és concretament. Cal tenir en compte que, com sempre, cal qüestionar els resultats oferts per les aplicacions, ja que com veiem en la mateixa imatge, altres resultats poden semblar molt similars, i no sempre el resultat ofert és el correcte.

4.3.3 – Publicitat i difusió

En aquest cas, la QBE s'utilitza per a la cerca de propostes per a l'obtenció de la imatge que realment es vol. És el camp on la precisió en la recuperació té menys incidència, ja que no hi ha un resultat amb un 100% de precisió, sinó que l'usuari tria la imatge més aproximada a la seva idea inicial, que no té per què ser la més definida i definitiva.

Un exemple de com es pot aplicar la recuperació d'imatges basada en el contingut dins de l'àmbit de la publicitat i difusió podria ser Hashtoc[2]. Hashtoc és un metacercador de productes de diferents botigues en línia que, a més de basar la seva cerca en hashtags, també permet l'ús d'imatges per a la cerca. És un entorn pensat de manera similar al de xarxes socials actualment populars com Twitter o Instagram. De fet, un altre exemple seria Pinterest que funciona de manera molt semblant a Hashtoc: mitjançant l'etiquetatge i cerca per imatges similars.

Pel que fa a la recuperació de logotips, la tesi de Robles Sanchez (2004) ja recull casos d'altres autors (Kwan et al., 2002) on s'utilitzen tècniques de recuperació de logotips a partir de la localització dels segments que conformen les seves vores. Més recentment, l'article Wu, Xiao i Hong (2018) descriu un nou enfocament del mateix concepte que podria ser d'aplicació transversal.

4.3.4 – Medicina

En aquest camp s'emmarcaria l'article de Pelka, Nensa i Friedrich (2018) on proposen l'anotació automatitzada d'imatges mitjançant l'*Image Retrieval of Medical Applications* (IRMA). Aquesta anotació automatitzada, junt amb tècniques de millora de la imatge, facilita la diagnosi de malalties assistida per ordinador (*Computer Aided Diagnosis* o CAD). Aquesta metodologia de treball facilita la localització de tumors en proves mèdiques, gràcies a la modificació d'elements com el color en la imatge.

També hi trobem altres exemples, com el presentat per Gupta, Kumar Dash i Mukhopadhyay (2016) que utilitzen factors avançats com la intensitat i el gradient de color i la seva direcció per a la detecció i diagnòstic d'alteracions pulmonars.

Recentment s'ha aplicat el reconeixement facial 3D en el camp sanitari que ha permès la detecció precoç de malalties rares durant la infància (El Español, 2019)[3]. Les tècniques utilitzades són similars a les utilitzades en la detecció de forma i vores, com descriuen Ponceleón i Slaney (2011) i Wu, Xiao i Hong (2018).

[2] Hashtoc <https://hashtoc.com> [Consulta: 20 maig 2019]

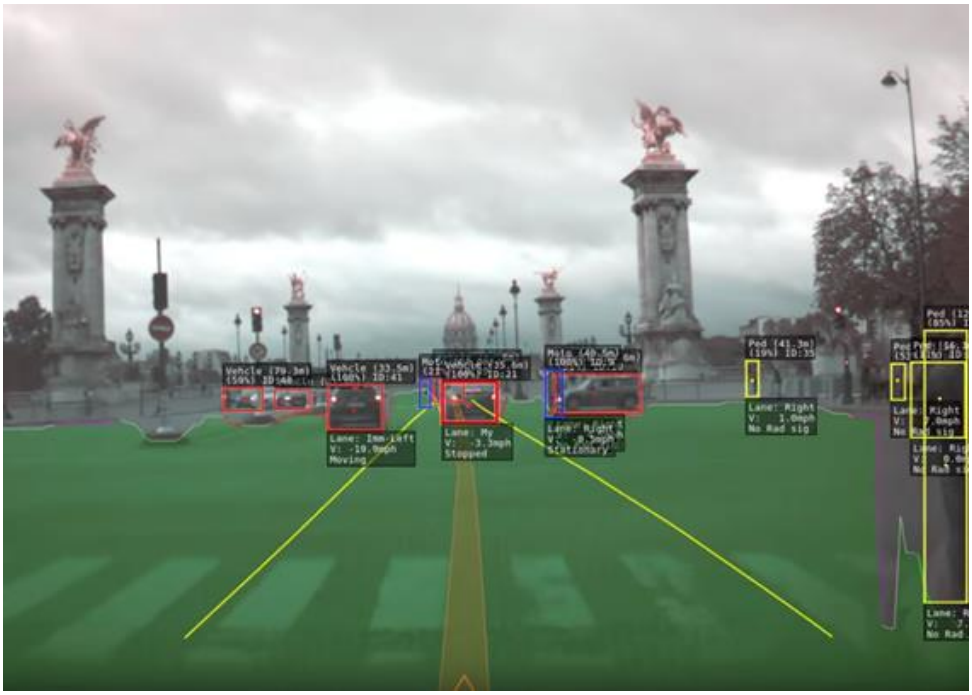
[3] Pel que fa al projecte, s'enllaça en la noticia al portal <https://cliniface.org> [Consulta: 20 maig 2019].

4.3.5 – Seguretat

Aquí s'hi poden trobar els elements de reconeixement biomètric (com el reconeixement facial a partir d'una imatge o d'un vídeo o el de petjades dactilars). Molts d'aquests elements estan ja presents en telèfons mòbils, que incorporen tècniques de reconeixement dactilar i facial com a elements de seguretat per protegir els continguts dels usuaris dels terminals i que actuen en alguns casos com a incentius de venda dels dispositius.

La premsa ja recull algunes aplicacions governamentals del reconeixement facial orientades a la seguretat i el control de la població. Per exemple, una notícia d'Albert Molins de 2018 a La Vanguardia explica com el govern xinès vol puntuar el comportament dels seus ciutadans a partir de l'anàlisi de les seves accions en els espais públics, que funcionaria mitjançant el reconeixement facial extret a partir de les imatges dels vídeos captats.

També es pot considerar la detecció de vehicles d'alguns models d'alta gamma com una aplicació del reconeixement d'imatges com a element de seguretat, com es pot apreciar en la Imatge 3. Aquesta característica és el pas previ per a la conducció autònoma de vehicles: detectant els vehicles i el càlcul de la seva velocitat i trajectòria dins de la via (elements que, com veurem, estan començant a aplicar-se en la recuperació de documents audiovisuals), el vehicle pot anticipar-se i respondre a l'estat de la circulació, millorant-ne la seguretat. Encara existeixen limitacions, com s'indica en l'article de Marcos Merino (2019) i fins ara la detecció de vehicles només és una ajuda o assistència a la conducció, però en un futur permetrà que el cotxe tingui un pilot completament automàtic i autònom i, per tant, no requerirà algú al volant.



Imatge 3: Aplicació del reconeixement d'imatges en la conducció semi-autònoma
En la imatge es poden veure els cossos identificats pel sistema "autopilot" de Tesla com a cotxes (en vermell), motos (en blau) i vianants (en groc) i les seves velocitats relatives (en milles per hora). Imatge extreta del vídeo <https://www.youtube.com/watch?v=_1MHGUC_BzQ> [Consulta: 20 Maig 2019], on es presenten tot mena de situacions.

4.4 – Situació actual i tendències

La imatge ha esdevingut un element especialment rellevant dins dels corrents tecnològics actuals, essent l'element clau en gran part de les xarxes socials més populars actualment. És per aquest motiu que la cerca i recuperació d'imatges basada en el seu contingut és un camp que pren major importància com a element de cerca i recuperació d'informació. La seva tendència és la d'esbrinar noves combinatòries i maneres d'analitzar l'aplicació dels diferents descriptors, i una redistribució del seu pes en el procés de recuperació de continguts, comprovant-ne l'impacte (Hao, Ge i Wang, 2018). Aquestes investigacions són rellevants pel que fa a no només a l'augment de l'eficàcia en la recuperació d'imatges, sinó en l'aplicació d'aquestes tècniques en entorns de vídeo, on els diferents punts claus del document el conformen precisament imatges.

S'han obert noves possibilitats en la manera de proporcionar una Consulta per Exemple (QBE) amb sistemes que, alternativament a la cerca mitjançant una imatge de referència, permeten les Consultes per Esbós (*Query By Sketch* o QBS), on l'usuari pot, proporcionant un dibuix, recuperar una imatge semblant a aquest dibuix (Gasser, Rossetto i Schuldt, 2019). Si bé els autors fan una distinció clara entre QBE i QBS, en realitat el QBS no deixa de ser un QBE amb menor definició de detalls, prescindint generalment de les textures i donant un major pes a les formes i colors. De fet, ja existeixen programaris capaços de recrear imatges per ordinador a partir d'un dibuix poc definit per procediments similars (Raya, 2019). Per altre banda, en alguns sistemes, especialment els de vídeo, la QBE consisteix en la captura de vídeo en temps real. Per exemple, en el cas de la conducció autònoma que hem vist en la Imatge 3.

A banda de la recuperació basada en el contingut, molts sistemes de recuperació han optat per l'anotació automatitzada a partir de la imatge (Pelka, Nensa i Friedrich, 2018) o per l'etiquetatge social (o *crowdsorcing*) (en el cas de l'àudio, Spina et Al., 2017, tot i que molts arxius també apliquen aquestes solucions d'etiquetatge social per les imatges) com a solucions complementàries.

Dins del camp dels sistemes aplicats, com hem vist en l'anterior apartat, destaca l'aplicació del reconeixement facial tant en els camps de medicina com de seguretat. Cal dir que en aquest sentit, existeixen certs aspectes ètics i de privacitat que el ràpid desenvolupament de la tecnologia fa que sovint quedin una mica deixats de banda, com destaca Enrique Dans (2019) en un article recent. Un exemple, de la necessitat de plantejar-se aspectes ètics i de seguretat de l'usuari d'aquests sistemes és la possibilitat de trampejar el sistema de pilot semiautomàtic d'un cotxe, fet que pot posar en perill la vida de l'usuari d'aquests sistemes (Merino, 2019).

En aquest sentit, ja es comencen a veure les primeres iniciatives per, si més no, establir certs límits a l'ús d'aquestes tecnologies allà on el seu ús comença a estar fortament desplegat. Per exemple, San Francisco ha decidit recentment proposar per a votació la prohibició del reconeixement facial sense autorització (Musil, 2019). Aquesta prohibició és especialment significativa perquè moltes de les empreses basades en noves tecnologies tenen la seva central operativa en aquesta ciutat.

5 – La recuperació d'àudio basada en el contingut (CBAR)

5.1 – Definició i conceptes bàsics

La recuperació d'àudio basada en el contingut (CBAR per les sigles de *Content-Based Audio Retrieval*) consisteix en l'obtenció d'un document sonor i/o la seva informació associada, a partir d'un fragment o la totalitat d'un primer document sonor o, fins i tot, la cerca de la seva expressió textual.

Dins de la recuperació d'àudio basada en el seu contingut cal distingir el que és el *Content-Based Music Retrival* (CBMR) o *Music Information Retrieval* (MIR) (Bognadov et al, 2013) del que és el reconeixement i recuperació del contingut basat en la veu (*Audio Speech Recognition* o ASR) (Mertens, Huang, Gotlieb et al., 2012; Spina et al., 2017), ja que hi apliquen necessitats d'informació i procediments de recuperació diferents en un cas i l'altre. Per una banda, en el cas de la CBMR o MIR, interessa la recuperació de la cançó o d'elements vinculats a les metadades de la cançó, als quals s'arriba per la comparació del QBE amb el contingut. Per l'altra, en el cas del ASR es busquen elements com la transcripció del contingut d'un registre d'àudio o la identificació dels interlocutors.

A diferència dels documents textuals i gràfics, els documents sonors no són directament interpretables: com indiquen Spina et al. (2017) "'visualitzar' un sumari significa reproduir un fragment d'àudio". A més, els documents sonors tenen un important buit semàntic, i requereixen la revisió del senyal d'àudio per extreure'n la informació. Ponceleón i Slaney (2011) expliquen que les ones de so són, en realitat, una simplificació d'un conjunt de sons captats i, que tot i amb aquesta simplificació, per tenir una mínima fidelitat, l'ona es modela 44.100 vegades per segon. Per tant, qualsevol recuperació d'informació basada en un contingut sonor requerirà una anàlisi prèvia del contingut, la intenció és automatitzar aquesta anàlisi.

5.2 – Elements descriptors que hi intervenen

Ponceleón i Slaney (2011) indiquen l'existència de tres elements clau en tot so:

- La **intensitat**, que és el volum del document sonor.

- El **to**, que és la posició del so dins d'una escala. Cada to té una **freqüència** auditiva única, i s'utilitza el *la* internacional (A3) com a referència perquè té un valor absolut de 440 Hz. Aquest element és especialment important, ja que serà el que es tractarà amb la transformada de Fourier que hem vist anteriorment.

- El **timbre**, que marca la diferenciació instrumental i el que permet la comprensió verbal. Tot i que el *la* internacional té la mateixa freqüència amb independència de l'instrument amb el qual es toqui, les seves harmòniques i reverberacions són diferents.

- Un quart element important que recull Wang (2006) però que Poncelón i Slaney (2011) no contemplen és la **duració** del so, probablement per simplificació i perquè es pot extreure de la **freqüència**.

De cara a la recuperació d'àudio, els més interessants són el to i el timbre. Cadascun d'aquests dos elements té diferents representacions gràfiques: els cromagrames pels tons i els espectrogrames pel timbre. Els cromagrames són la representació gràfica de la tonalitat d'un so (com es pot veure en la Figura 1 amb la representació d'una octava). Pel que fa als espectrogrames són la combinació de diferents cromagrames de múltiples octaves, i amb elles, la representació gràfica de la freqüència d'un so (Ponceleón i Slaney, 2011). Aquestes representacions gràfiques permeten establir i agrupar seqüències matemàtiques - conegudes com a vectors - interpretables per un ordinador, i per tant, es poden incloure en alguns algoritmes com els que s'han presentat anteriorment.

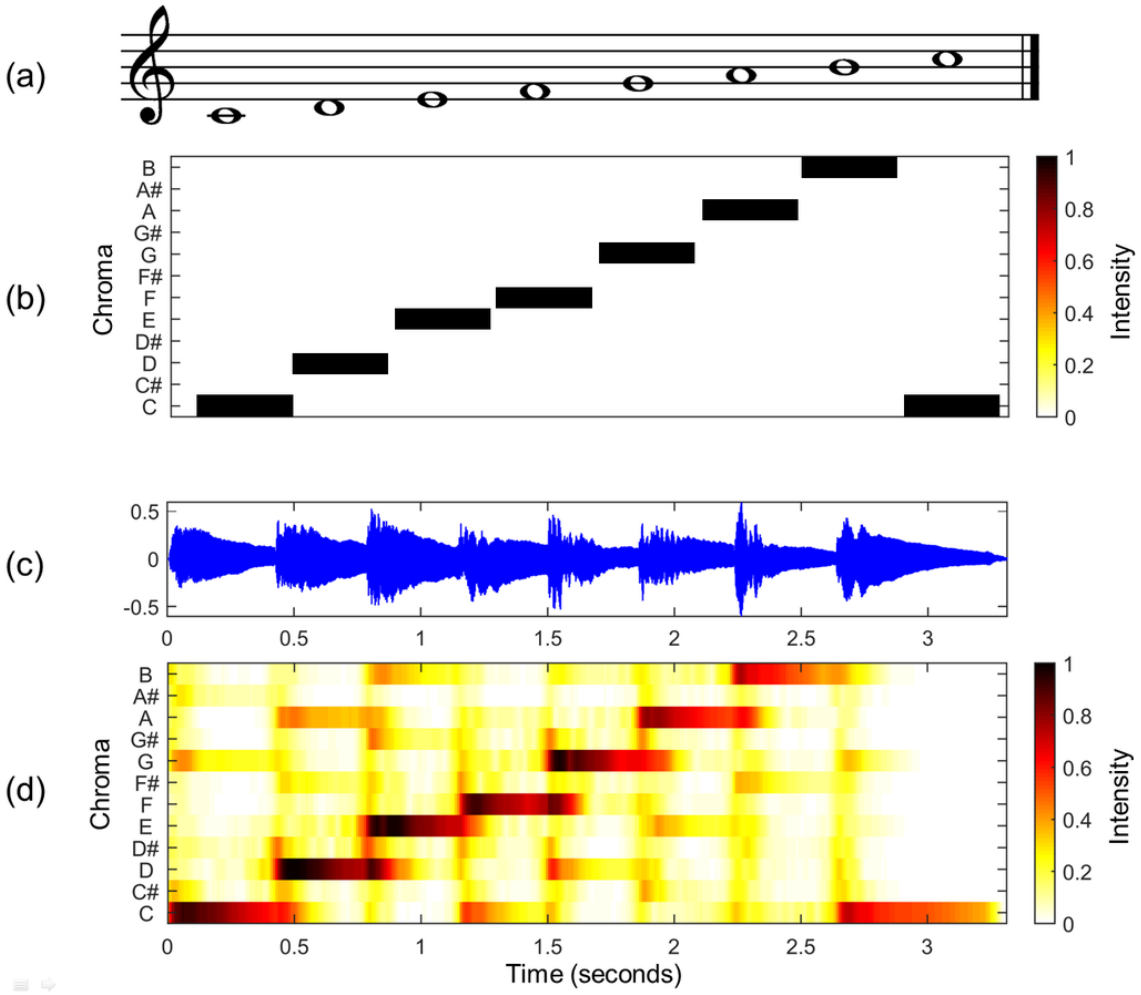


Figura 1: Cromagrama teòric (b) a partir d'una escala musical i real (d) de la mateixa escala en un document sonor
Imatge per Meinard Mueller, sota llicència CC BY-SA 3.0
(https://commons.wikimedia.org/w/index.php?curid=47816462)

5.3 – Aplicacions

5.3.1 – Identificació musical (Fingerprinting)

La identificació musical o *fingerprinting* consisteix en, donat un fragment d'àudio, cercar una coincidència perfecta en el seu espectrograma amb un altre document sonor dins d'una base de dades de documents sonors. Això és exactament el que fa, entre altres aplicacions, l'aplicació *Shazam*: grava la cançó que l'usuari vol identificar, la cerca en la seva base de dades i retorna tant la informació associada a la cançó com una mostra sonora.

El cas de *Shazam* és el més popular, i s'ha triat per ser un cas conegut i d'èxit, encara que hi ha altres aplicacions similars. (Per exemple, *Bing Music* -integrat amb l'assistent de veu Cortana-, *Google Sound Identification* o *SoundHound*). Per aquest motiu, és especialment interessant un article del mateix creador del servei, Avery Wang (2006) parlant de les funcionalitats del servei en els seus orígens. En aquest article, Wang (2006) planteja tres grans problemàtiques a l'hora de recollir la QBE: el soroll de fons, les interferències i la gestió de la base de dades. Sobre el soroll, Wang deixa en entreveure amb els exemples que es presenten que hi ha un conjunt més o menys previst de circumstàncies en les quals es captura el so. Pel que fa a les interferències, fa referència a la qualitat del micròfon que captura la mostra i alguns dels elements de millora del so dels dispositius que perjudiquen el reconeixement dels sons de fons, on sovint es troba la QBE. Per últim, pel que fa a la gestió de la base de dades, Wang diu que aquesta ha de gestionar les empremtes (*fingerprints*) del so. Aquestes empremtes, realitzades a partir dels punts més rellevants de l'espectograma, han de ser prou representatives però alhora prou compactes per a ser les més petites possibles, per poder respondre de manera àgil a les consultes rebudes pel servidor. Per tant, les QBE obtingudes no es comparen mai amb una cançó, sinó amb una representació matemàtica: passant de valors que ocuparien Megabytes en la base de dades a registres d'uns pocs Kilobytes.

Una altra aplicació, com indiquen Ponceleón i Slaney (2011), és la de la cerca d'inclusió de materials il·legals o dels que no se n'és el propietari dels drets. Un cas pràctic d'aquesta aplicació, seria el *ContentID* de Youtube (Popper, 2016), que utilitza *fingerprinting* aplicat al vídeo per detectar l'ús no autoritzat de cançons en la banda sonora dels vídeos i desmonetitzar o eliminar aquells vídeos amb continguts de tercers.

5.3.2 – Reconeixement i interpretació de la veu

El reconeixement de veu és el procés pel qual s'interpreta les paraules contingudes en un document sonor o audiovisual (Slaney i Ponceleón, 2011). Com en el cas del *fingerprinting,* el reconeixement de veu requereix l'eliminació del soroll de fons. En aquest cas, a més, l'algoritme cal que conegui l'idioma i el nombre de paraules que cal que reconegui.

Existeixen dos grans camps d'estudi:

- Per una banda, el del **reconeixement i recuperació dels continguts del discurs**. En aquest sentit, la cerca per veu de Google existeix des de fa bastant temps, però ha pres rellevància real amb la popularització dels assistents virtuals com poden ser Siri, Cortana o Alexa que poden donar resposta sintetitzada en veu. Aquests són assistents de veu que tradueixen ordres de veu a diferents *outputs*, podent obtenir resultats basats en continguts textuals, continguts no textuals i, fins i tot, la traducció d'aquestes ordres a senyals binàries, fent-los especialment interessants per aplicacions domòtiques.

- Per una altra banda, el de **reconeixement de l'interlocutor**, on no es té en compte el contingut del discurs, sinó la identitat de qui el fa. És possible reconèixer a una persona a partir de la seva veu que es coneix com a Passaport Vocal (Romero, 1999)[4] de les que es pot extreure elements com el sexe, una franja d'edat, la localització o el nivell socioeconòmic. Aquest passaport vocal pot actuar com a indicador biomètric, ja que tot i que hi ha veus similars, cada veu és única. Novament els elements clau per la identificació de la veu d'una persona són el to i el timbre.

[4] També es fa referencia al terme passaport vocal en l'article de divulgació de Ortiz, Ana María. "Criminales a los que la Policía caza por su voz". *El Mundo*. <https://www.elmundo.es/espana/2019/04/20/5cb9fdeafc6c83411f8b45bb.html?cid=SIN12201> [Consulta: 20 maig 2019]

5.4 – Situació actual i tendències

Molta de la investigació actual està enfocada a simplificar i donar un significat semàntic a l'àudio a partir del seu contingut, amb independència de si el contingut és música o un enregistrament de veu, per bé que és en aquest segon cas on es treu major partit d'aquestes capacitats. Per exemple, Bognadov et al. (2013) a través de l'anotació automatitzada i la generació d'una imatge representativa i Mertens, Huang, Gottlieb et al. (2012) a través de resums i sintetitzats del contingut dels documents sonors, de la mateixa manera que ja es feia en la recuperació de documents audiovisuals. Spina et al. (2017) apliquen de la mateixa manera aquests resums en entorns més propers a l'entreteniment, fent resums auditius de *podcasts* mitjançant el reconeixement automatitzat del discurs (ASR). Aquests resums són esbiaixats pensant en la posterior recuperació d'aquest contingut: si bé per ells mateixos no són representatius de la totalitat del contingut, s'entén que permet la recuperació d'aquells fragments més significatius.

En el cas que assenyalen Spina et Al. (2017), citant altres autors, el reconeixement automatitzat del discurs millora la precisió en la recuperació en un entorn -el de producció de *podcasts*- on els continguts ja tenen moltes metadades textuals introduïdes i estan bastant identificats. Per tant, l'ús de sistemes automatitzats de recuperació d'àudio, tot i que tendeixen a ser per si sols poc precisos, contribueixen positivament a la recuperació inclús en entorns ja etiquetats amb metadades. Un altre element interessant és el de la correcció de les transcripcions mitjançant el *crowdsorcing* i, amb aquestes correccions, la possibilitat d'explotació de les *Suport Vector Machines*. D'aquesta manera, s'aprofita el *machine learning* perquè la màquina es pugui autocorregir en transcripcions futures i així millorar progressivament la precisió en la transcripció.

Les grans empreses tecnològiques (Amazon, Google o Microsoft entre d'altres) ofereixen aplicacions que permeten la transcripció de grans quantitats d'àudio (García Nieto, 2019). A diferència de les aplicacions basades en el *fingerprinting*, no són encara gaire populars. L'article també assenyala que els motius, més que econòmics – és cert que no són aplicacions gratuïtes, però si força assequibles-, són de complicacions en l'ús de les aplicacions per part de l'usuari. Per exemple, en el cas de la solució de Google, es requereixen coneixements de programació, fet que allunya a alguns dels col·lectius que podrien ser més beneficiats per l'ús d'aquests sistemes, com poden ser els periodistes en el camp de la transcripció d'entrevistes o els estudiants pel que fa a la transcripció de classes magistrals.

Òbviament, hi ha aplicacions a la inversa: Aplicacions que són capaces de generar so a partir d'un text i un Passaport Vocal. No les tractarem amb major profunditat perquè no és l'objectiu d'aquesta investigació, però perquè quedi constància de la seva importància, ja s'està comparant a *project VoCo* d'*Adobe*[5] com el *Photoshop* del so. Al respecte hi ha una important

[5] Demostració de Project VoCo: Adobe. " #VoCo Adobe MAX 2016 (Sneak Peeks) | Adobe Creative Cloud". *Youtube* <https://youtu.be/I3l4XLZ59iw> [Consulta: 20 maig 2019]

preocupació sobre el fenomen conegut com a *Deep Fake* (Generació d'àudios o inclús vídeos d'elements que mai han succeït, però que són molt difícils de contrastar) que pot generar.

Pel que fa a la identificació musical, el seu èxit depèn en gran mesura de la qualitat del mostreig previ (en altres paraules, la descripció dels documents sonors a recuperar) i la mida de la base de dades del servei, ja que en línies generals el reconeixement és força bo. De fet, Wang (2006) preveia com una possible QBE l'ús de les *Query By Humming* (cerca per taral·lejat), i si bé algunes competidores com *SoundHound* si l'han implementat, no està clar que *Shazam* l'hagi incorporat.

Com en el cas de les imatges i el reconeixement facial, també s'especula molt amb les possibles implicacions ètiques i de privacitat relacionades amb l'ús de sistemes de recuperació d'àudio basats en el contingut. Els assistents de veu per funcionar correctament, han d'estar constantment a l'escolta de les comandes que activen el seu funcionament. Aquest fet implica que per força, la intimitat dels seus usuaris es pugui veure compromesa, com demostra per exemple una notícia referent a *Alexa* en la que es deia que els treballadors d'*Amazon* poden escoltar el que capta l'assistent de veu (Day, Turner i Drozdiak, 2019).

6 – La recuperació de documents audiovisuals en relació al MMIR

6.1 – Conceptes bàsics

Els documents audiovisuals són el resultat d'una combinació d'un conjunt de documents gràfics i sonors. Per aquest motiu, les tècniques de recuperació basades en el contingut emprades en la recuperació d'imatges o sons, que hem detallat anteriorment, també s'apliquen als vídeos. Sovint, el més senzill és que la recuperació d'aquests documents es faci mitjançant aquestes tècniques, ja que cal tenir en compte que els documents audiovisuals són molt pesants en termes computacionals, tant pel que fa al seu ús com a element de cerca com per a la seva recuperació, que a més requereixen reproductors compatibles amb la seva codificació per a la seva representació.

Elements propis d'aquests documents, com les trajectòries dels moviments i la mateixa combinació d'imatge-so, doten als vídeos d'un major context, que permeten reduir el buit semàntic (Mertens, Huang, Gottlieb et Al., 2012). Al respecte, Ponceleón i Slaney (2011) afegeixen que "les persones són més eficients revisant material visual que d'àudio" i, de fet, quan parlaven del buit semàntic, els autors situaven la música com l'element amb el buit semàntic més gran. El motiu és la contextualització que assenyalen Mertens, Huang, Gotlieb et. Al (2012) en el seu article.

6.1.1 – Anàlisis de Contingut del Vídeo (MoCA)

Ponceleón i Slaney (2011) indiquen que la indexació dels vídeos es fa partint de resums de vídeo. Aquests resums no només són útils per la seva indexació, sinó que una vegada ofert com a resultat, el resum facilita la representació dels seus continguts de manera més fidel que un resum textual, amb el que el buit semàntic hauria de ser inferior. Per elaborar els resums, els autors identifiquen 4 passos clau:

1. **Anàlisi i divisió del vídeo**: cal segmentar el vídeo en fragments més petits per tal de gestionar-los. Una bona referència a l'hora de segmentar un vídeo és aprofitar les transicions entre escenes.

2. **Classificació segons tipologia dels continguts (imatge, àudio, text):** a partir dels diferents continguts dels fragments es pot establir un primer criteri de rellevància.

3. **Selecció dels segments més representatius:** amb els elements classificats del primer criteri de rellevància, es fa una segona anàlisi per seleccionar els fragments més rellevants i representatius del vídeo.

4. **Generació de la visualització, que pot ser de dos tipus:**

    - **Estàtica**: a partir d'imatges concretes del metratge. Poden ser impresos en paper i poden ser resums textuals o fets amb keyframes o fotogrames claus. Solen incloure comentaris i marques de temps. En la selecció de keyframes regeixen criteris vistos anteriorment en l'apartat de recuperació d'imatges basades en el seu contingut (els keyframes són escollits per criteris com el color, la textura i el moviment d'aquests).

    - **Dinàmica**: a partir dels fragments de vídeo i/o so detectats com a rellevants. Aquests sumaris poden ser diapositives amb controls de vídeo, un Storyboard en moviment o tràilers.

Teòricament, el MoCA no només permet la indexació i millora de la representativitat dels continguts, sinó també la recuperació de vídeos basada en el contingut, com proposen Barrios i Bustos (2011) per a la identificació de còpies. Però cal tenir en compte que és força més senzill i efectiu cercar a partir d'imatges i sons continguts en el vídeo que no pas amb el vídeo pròpiament, pel seu propi pes i processament computacional.

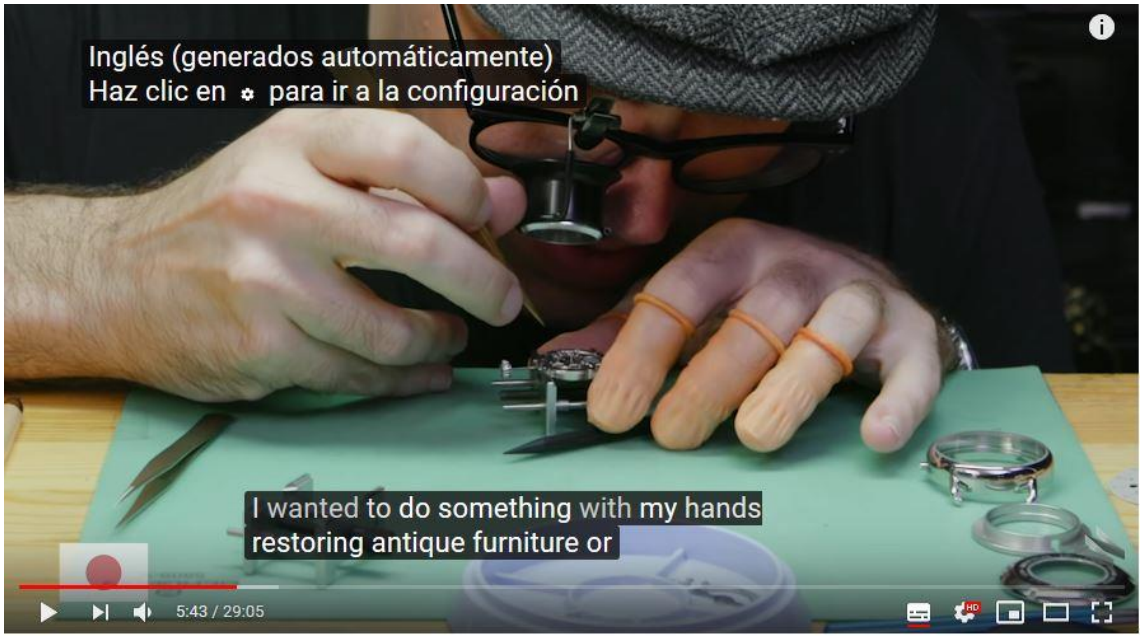
6.2 – Tècniques i aplicacions

6.2.1 – Predicció de moviments

Indiquen Slaney i Ponceleón (2011) que la predicció de moviments treballa d'una manera similar a la recuperació d'imatges basada en el contingut, i en concret a la detecció de textures, vores i posició dins de la imatge: mitjançant la similitud entre grups de píxels propers que conformen un objecte visual. D'aquests grups detectats, se'n fa un seguiment fotograma a fotograma i, un cop establerts en un primer fotograma i el successiu, es poden establir vectors de moviment que permeten el seguiment d'aquests objectes. Com hem vist en alguns exemples anteriorment (un en concret, el que es pot veure en la Imatge 3), és el pas previ a la recuperació d'imatges basada en el contingut: és més fàcil detectar en una imatge concreta un determinat objecte i seguir-ne el moviment que anar identificant-lo a cada fotograma.

Cal tenir en compte que, per simplificar-ho, tant en aquesta explicació com en les corresponents a les imatges, no s'han tingut en compte les compressions dels diferents formats d'imatge i vídeo. Aquests procediments utilitzen algoritmes similars per agrupar píxels segons el color i fer més lleugeres les imatges (i en el cas del vídeo, cadascun dels fotogrames que conformen), i per tant, alteren el funcionament de la predicció de moviments, que a major compressió d'imatges, major complexitat i menor precisió en la detecció.

6.2.2 – Audiovisual Speech Recognition (AVSR)



Imatge 4: Generació de subtítols automàtics a partir d'AVSR a YouTube
Exemple extret del video <https://youtu.be/7hSZFO5QUmE?t=343> [Consulta: 20 Maig 2019]

L'Audiovisual Speech Recognition (AVSR) és l'aplicació en l'entorn audiovisual de l'Automated Speech Recongnition (ASR) que Spina et Al. (2017) detallen en el seu article i que ja hem comentat anteriorment en el cas del so. Si en el cas de Spina et al. aquest reconeixement podia arribar a permetre la indexació dels *podcasts* a partir del seu contingut, no és difícil de pensar que, en el cas dels documents audiovisuals poden permetre elements com el subtitulat automàtic per millorar la comprensió del document. Un exemple és el que es mostra amb la Imatge 4: opcionalment els creadors de contingut poden afegir subtítols en diversos idiomes o permetre que Youtube els generi automàticament mitjançant aquesta tecnologia.

Ponceleón i Slaney (2011) expliquen que com a pas previ cal unificar el vídeo i el so, però que es pot aplicar els Models Ocults de Markov abans o després de la unió del vídeo i el so. Indiquen també que el millor seria aplicar l'algoritme després de la fusió audiovisual o apostar per una solució hibrida, fet que corroboraria l'afirmació que feien amb anterioritat sobre el buit semàntic.

7 – Conclusions i possibles línies d'investigació futures

Des del punt de vista de la informació, la cerca i recuperació d'informació no textual basada en el seu contingut hauria de permetre no només donar solució al creixent ritme de generació de continguts, sinó una major objectivitat en el procés de descripció de la informació i, en conseqüència, resultats més acurats i precisos en la recuperació documental dins de grans conjunts d'informació. A la llarga, les metadades han de passar de ser un element de descripció mínima a un complement dels sistemes de recuperació d'informació no textuals. Cal dir que l'existència d'aquests sistemes de recuperació no implica que s'hagin de deixar de descriure aquest tipus de documents, sinó que aporta una solució temporal per a poder recuperar-los fins que es decideixi la conveniència o no de descriure'ls.

Amb tot, l'aplicació de sistemes de recuperació multimèdia basats en el seu contingut no exclou l'ús de metadades, sinó que busca la seva generació i anotació automàtiques i les utilitza com a elements contextuals per reduir el buit semàntic. Aquest camp té, per tant, més interès en aquelles tipologies amb major buit semàntic. El *crowdsorcing* i altres eines contextuals, com l'elaboració de perfils d'usuari, juguen un paper important tant en l'etiquetatge amb metadades manual com amb l'aprenentatge de les *Suport Vector Machines* que han de permetre fer les anotacions i transcripcions automatitzades. L'obtenció de retroacció de l'usuari millora la rellevància i la preferència (Spina et al., 2017). Cal tenir en compte les implicacions i limitacions relacionades.

Molts dels algoritmes emprats per a la recuperació d'informació multimèdia basada en el seu contingut són compartits entre diferents tipologies de contingut. Aquest fet no ha de resultar estrany: Tant el color, com el so i les trajectòries com a elements més representatius de cadascuna de les diferents tipologies documentals vistes en aquest treball són, al nivell més bàsic d'abstracció, senyals calculables mitjançant fórmules matemàtiques extretes a partir dels fenòmens físics. En aquest treball, hem vist com es poden descriure mitjançant les transformades de Fourier i de Wavelets i com es poden cercar posteriorment amb els models ocults de Markov i mixtos gaussians, que s'apliquen tant en entorns CBIR (en especial, en entorns amb textures dinàmiques com els que recullen Paygude i Vyas (2018)) com en entorns CBAR (com s'indica en l'article de Bognadov et al. (2013)). El contingut es tradueix en màquines de suport de vectors que aprenen amb models estadístics abstractes i transversals, per aquest motiu, no és estrany que existeixin alguns casos en els quals s'han reunit en un únic sistema de recuperació (Gasser, Rosseto i Schuldt, 2019), encara que segons indiquen ells mateixos, "Encara hi ha pocs casos de solucions integrades". En el cas dels entorns CBAR, la complexitat més gran del format implica altres algoritmes propis tant pel que fa a la seva representació visual com pel que fa al procés d'indexació i recuperació, previs a l'aplicació dels models fusionats entre Markov i mixtos gaussians.

Els algoritmes i tècniques presentats esdevenen components molt importants en el *Machine Learning* i l'ús de les xarxes neuronals i la intel·ligència artificial per a l'automatització de tasques. Malgrat tot, aspectes com les implicacions ètiques de l'ús, abús i trampejat d'aquests

sistemes haurien de ser, amb certa urgència, línies d'investigació futures. En aquest treball s'ha exposat com funcionen internament aquests sistemes i quines són les seves aplicacions, però només s'intueixen les seves implicacions pel que fa a elements com la privacitat, l'ús de les dades amb altres finalitats que les inicialment anunciades (generalment comercials) o inclús la integritat física dels usuaris.

8 – Bibliografia

Publicacions i articles científics

AGUSTÍ, M., MIGUEL, J., GONZÁLEZ, V. i ROCAMORA, M.C., 2003. Recuperación por contenido en bases de datos de imágenes basada en wavelets : aplicación al diseño del textil. , no. Desembre 2014.

BARRIOS, J.M. i BUSTOS, B., 2011. P-VCD: A pivot-based approach for Content-Based Video Copy Detection. 2011 IEEE International Conference on Multimedia and Expo [en línia]. S.l.: IEEE, pp. 1-6. [Consulta: 20 Maig 2019]. ISBN 978-1-61284-348-3. DOI 10.1109/ICME.2011.6012212. Disponible a: <http://ieeexplore.ieee.org/document/6012212/>.

BOGDANOV, D., HARO, M., FUHRMANN, F., XAMBÓ, A., GÓMEZ, E. i HERRERA, P., 2013. Semantic audio content-based music recommendation and visualization based on user preference examples. Information Processing and Management [en línia], vol. 49, no. 1, pp. 13-33. ISSN 03064573. DOI 10.1016/j.ipm.2012.06.004. Disponible a: <http://dx.doi.org/10.1016/j.ipm.2012.06.004>.

GASSER, RALPH; ROSSETO, LUCA; SCHULDT, H., 2019. Towards an All-Purpose, Content-Based Multimedia Information Retrieval System. ,

GUPTA, R. Das, DASH, J.K. i MUKHOPADHYAY, S., 2017. Content based retrieval of interstitial lung disease patterns using spatial distribution of intensity, gradient magnitude and gradient direction. 2016 International Conference on Systems in Medicine and Biology, ICSMB 2016, vol. 1, no. January, pp. 58-61. DOI 10.1109/ICSMB.2016.7915087.

HAO, Z., GE, H. i WANG, L., 2018. Visual attention mechanism and support vector machine based automatic image annotation, PLoS ONE [en línia], vol. 13, no. 11. ISSN 19326203. DOI 10.1371/journal.pone.0206971. Disponible a: <http://dx.doi.org/10.1371/journal.pone.0206971>.

MERTENS, R., HUANG, P.-S., GOTTLIEB, L., FRIEDLAND, G., DIVAKARAN, A. i HASEGAWA-JOHNSON, M., 2012. On the Applicability of Speaker Diarization to Audio Indexing of Non-Speech and Mixed Non-Speech/Speech Video Soundtracks, International Journal of Multimedia Data Engineering and Management [en línia], vol. 3, no. 3, pp. 1-19. ISSN 1947-8534, 1947-8534. DOI http://dx.doi.org/10.4018/jmdem.2012070101. Disponible a: <http://search.proquest.com/docview/1531921562?accountid=142596>.

MORENO-SCHNEIDER, J., MARTÍNEZ, P. i MARTÍNEZ-FERNÁNDEZ, J.L., 2017. Combining heterogeneous sources in an interactive multimedia content retrieval model. Expert Systems with Applications, vol. 69, pp. 201-213. ISSN 09574174. DOI 10.1016/j.eswa.2016.10.049.

PAYGUDE, S. i VYAS, V., 2018. Dynamic Texture Segmentation Approaches for Natural and Manmade Cases: Survey and Experimentation. Archives of Computational Methods in Engineering [en línia], no. 0123456789. ISSN 1134-3060. DOI 10.1007/s11831-018-09305-9. Disponible a: <http://link.springer.com/10.1007/s11831-018-09305-9>.

PEDREGOSA, F. et al., 2019. Gaussian mixture models. Journal of Machine Learning Research [en línia], 12, p. 2825-2830, 2011. [Consulta: 20 Maig 2019]. Disponible a: <https://scikit-learn.org/stable/modules/mixture.html>.

PELKA, O., NENSA, F. i FRIEDRICH, C.M., 2018. Annotation of enhanced radiographs for medical image retrieval with deep convolutional neural networks, Plos One [en línia], vol. 13, no. 11, pp. e0206229. ISSN 1932-6203. DOI 10.1371/journal.pone.0206229. Disponible a: <http://dx.plos.org/10.1371/journal.pone.0206229>.

PONCELEÓN, D. i SLANEY, M., 2011. Multimedia Information Retrieval. Modern information retrieval: the concepts and technology behind search. 2nd. Harlow [etc.] : Addison-Wesley / Pearson, pp. 48-6950-48-6950. ISBN 9780321416919

RAMEZANI, M. i YAGHMAEE, F., 2016. A review on human action analysis in videos for retrieval applications. Artificial Intelligence Review, vol. 46, no. 4, pp. 485-514. ISSN 15737462. DOI 10.1007/s10462-016-9473-y.

ROMERO, C.D., 1999. The Vocal Passport, a technique to outline profiles of criminals from their voice recordings. Proceedings IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology (Cat. No.99CH36303) [en línia]. S.l.: IEEE, pp. 77-79. [Consulta: 20 Maig 2019]. ISBN 0-7803-5247-5. DOI 10.1109/CCST.1999.797896. Disponible a: <http://ieeexplore.ieee.org/document/797896/>.

SÁNCHEZ, Ó.D.R., 2004. Técnicas de recuperación por contenido para imagen y vídeo en arquitecturas paralelas [en línia]. S.l.: Universidad Politécnica de Madrid. Disponible a: <http://oa.upm.es/182/>.

SPINA, D., TRIPPAS, J.R., CAVEDON, L. i SANDERSON, M., 2017. Extracting audio summaries to support effective spoken document search. Journal of the Association for Information Science and Technology [en línia], vol. 68, no. 9, pp. 2101-2115. [Consulta: 20 maig 2019]. ISSN 23301635. DOI 10.1002/asi.23831. Disponible a: http://doi.wiley.com/10.1002/asi.23831.

VALLEZ, MARI; PEDRAZA-JIMENEZ, R., 2007. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual i áreas afines. Hipertext.net [en línia], no. 5. [Consulta: 20 maig 2019]. ISSN 1695-5498. Disponible a: <https://www.upf.edu/hipertextnet/numero-5/pln.html#problematica-procesamiento-lenguaje-natural>.

WANG, A., 2006. The Shazam music recognition service. Communications of the ACM [en línia], vol. 49, no. 8, pp. 44. [Consulta: 20 maig 2019]. ISSN 00010782. DOI 10.1145/1145287.1145312. Disponible a: <http://mendeley.csuc.cat/fitxers/c11909a14e991cc5212b1c1037aadcfc>.

WU, M., XIAO, W. i HONG, Z., 2018. Similar image retrieval in large-scale trademark databases based on regional and boundary fusion feature, Plos One [en línea], vol. 13, no. 11, pp. e0205002. ISSN 1932-6203. DOI 10.1371/journal.pone.0205002. Disponible en: <http://dx.plos.org/10.1371/journal.pone.0205002>.

Articles de divulgació

BARREDO, Á., 2019. Cuando tu cara lo dice todo: el fantástico i terrorífico futuro del reconocimiento facial. La Vanguardia [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.lavanguardia.com/tecnologia/20190422/461771463993/reconocimiento-facial-futuro.html>.

CONDLIFFE, J., 2019. "Digital music couldn't exist without the fourier transformation". Gizmodo [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://gizmodo.com/digital-music-couldnt-exist-without-the-fourier-transfo-1699155287>.

DANS, E., 2019. Reconocimiento facial: cuando la tecnología avanza demasiado rápido. enriquedans.com [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.enriquedans.com/2019/04/reconocimiento-facial-cuando-la-tecnologia-avanza-demasiado-rapido.html>.

DAY, MATT; TURNER, G.D.N., 2019. Is Anyone Listening to You on Alexa? A Global Team Reviews Audio. Bloomberg [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio>

EL ESPAÑOL, 2019. Usan el reconocimiento facial en 3D para detectar enfermedades raras en niños. Omicrono El Español [en línia], 2019. [Consulta: 20 maig 2019]. Disponible a: <https://omicrono.elespanol.com/2019/04/reconocimiento-facial-en-3d-detectar-enfermedades/>.

GARCÍA NIETO, J., 2019. El estado del arte del software para transcribir entrevistas, clases magistrales i juicios. Xataka [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.xataka.com/especiales/estado-arte-software-para-transcribir-entrevistas-clases-magistrales-juicios>.

MERINO, M., 2019. Unas pegatinas en el asfalto bastan para «hackear» el piloto automático de un Tesla… i convencerle para ir en dirección contraria. Xataka [en línea]. [Consulta: 20 maig 2019]. Disponible a: <https://www.xataka.com/inteligencia-artificial/unas-pegatinas-asfalto-bastan-para-hackear-piloto-automatico-tesla-convencerle-para-ir-direccion-contraria>.

NIETO, A., 2019. Alguien ha hecho el video perfecto para todos los que sufrimos intentando entender la trasnformada de Fourier. Xataka [en línea]. [Consulta: 20 maig 2019]. Disponible a: <https://www.xataka.com/otros/alguien-ha-hecho-el-video-perfecto-para-todos-los-que-sufrimos-intentando-entender-la-transformada-de-fourier>.

MOLINS RENTER, A., 2018. China estrena su 'Gran Hermano'. La Vanguardia [en línia]. Barcelona, 3 mayo 2018. [Consulta: 20 maig 2019]. Disponible a: <https://www.lavanguardia.com/internacional/20180503/443196686690/china-puntuacion-ciudadanos-delitos-sociales.html>.

MUSIL, S., 2019. La prohibición del reconocimiento facial en San Francisco, lista para votación. Cnet en español [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.cnet.com/es/noticias/prohibicion-reconocimiento-facial-san-francisco-votacion/>.

ORTIZ, A.M., 2019. Criminales a los que la Policía caza por su voz. El Mundo [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.elmundo.es/espana/2019/04/20/5cb9fdeafc6c83411f8b45bb.html?cid=SIN12201>.

POPPER, B., 2016. YouTube to the music industry: here's the money. The Verge [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://www.theverge.com/2016/7/13/12165194/youtube-content-id-2-billion-paid>.

RAYA, A., 2019. Este programa de Nvidia convierte dibujos de Paint en fotografías ultrarrealistas. Omicrono El Español [en línia]. [Consulta: 20 maig 2019]. Disponible a: <https://omicrono.elespanol.com/2019/03/fotos-a-partir-de-dibujos/>.