UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

MÈTODES DEEP LEARNING PER L'ANOTACIÓ DE PERSONES EN SEQUÈNCIES DE VÍDEO

Treball de fi de grau presentat a la

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

per

Oriol Vila Clarà

En compliment parcial dels requeriments per al grau en

CIENCIES I TECNOLOGIES DE LES TELECOMNICACIONS

Tutor: Josep Ramon Morros

Barcelona, June 2016

Abstract

In the recent years, the demand for unsupervised annotation tools to annotate and classify large audiovisual datasets has grown considerably. One of these tasks is concretely addressed on TV broadcast videos, to determine who and when appears in a video sequence.

This work is aimed on exploring deep learning methods for face feature extraction and the implementation of a verification system in order to boost the performance of person recognition tasks.

A comparison between different identification methods that have been developed during this project is made, with the aim of evaluating their performance and to conclude which ones are the best.

Finally, a comparison between the results obtained with this system and the one proposed by the 2015 UPC System Mediaeval Multimodal Person in Discovery TV Broadcast is done. Achieving a performance boost up to 5.8% in terms of Mean Average Precision.

Resum

En els darrers anys ha aparegut la necessitat de disposar d'eines d'anotació no supervisada per tal de classificar i anotar grans conjunts de dades audiovisuals. Una d'aquestes tasques recau en anotar seqüencies de vídeo de TV per tal de determinar qui i quan apareix en un vídeo.

Aquest treball es centra en explorar sistemes d'extracció de característiques basats en tècniques de *deep learning* i la implementació d'un sistema de verificació per tal de millorar aquestes tasques d'anotació.

En el transcurs del projecte es realitza una comparativa de les diferents metodologies d'anotació que s'han desenvolupat amb l'objectiu d'avaluar-ne el rendiment i acabar seleccionant-ne les millors

Finalment s'acaba comparant els resultats del sistema final amb la proposta de UPC System for the 2015 MediaEval Multimodal Person Discovery in Broadcast TV task, obtenint una millora superior al 5,8% en termes de Mean Average Precision.

Resumen

En los últimos años ha aparecido la necesidad de disponer de herramientas de anotación no supervisada para clasificar y anotar grandes conjuntos de datos audiovisuales. Una de estas tareas recae en anotar secuencias de video de TV para determinar quién y cuando aparece en un vídeo.

Este trabajo se centra en explorar sistemas de extracción de características basados en técnicas de *deep learning* i la implementación de un sistema de verificación para mejorar estas tareas de anotación.

En el transcurso del proyecto se realiza una comparativa de las diferentes metodologías de anotación que se han desarrollado con el objetivo de evaluar su rendimiento y acabar seleccionando las mejores

Finalmente se termina comparando los resultados del sistema final con la propuesta de UPC System for the 2015 Mediaeval Multimodal Person Discovery in Broadcast TV task obteniendo una mejora superior al 5,8% en términos de Mean Average Precision.

Agraïments

En primer lloc agrair la ajuda i consell proporcionat per el meu tutor de projecte Josep Ramon Morros. També agrair la col·laboració de l'Albert Gil per el seu suport tècnic en els problemes relacionats amb els servidors de l'escola i a tots els professors en general que han aconseguit transmetre'm la passió per el món de la ciència i tecnologia.

Finalment agrair també a la meva família per el seu enorme suport moral i financer durant tots aquests anys.

Moltes Gràcies a tots.

Historial de revisió i aprovació

Revision	Date	Purpose
0	26/05/2016	Document creation
1	25/06/2016	Document revision
2	27/06/2016	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Oriol Vila Clarà	vlcl_ori_13@hotmail.com
Josep Ramon Morros	ramon.morros@upc.edu

Written by: Oriol Vila Clarà		Reviewed and approved by: Josep Ramon Morros	
Date	26/05/2016	Date	27/06/2016
Name	Oriol Vila Clarà	Name	Josep Ramon Morros
Position	Project Author	Position	Project Supervisor

Taula de continguts

Llista de Figures

Llista de Taules:

1. Introducció

La enorme quantitat de dades visuals (vídeos) que es genera actualment crea una forta necessitat d'eines d'anotació que facin possible la cerca i recuperació de la informació present en els vídeos. Una de les informacions més rellevants és la identitat de les persones. En aquest context, l'anotació consisteix en determinar qui apareix i en quins instants.

Per tal de realitzar aquestes tasques d'anotació no supervisada s'utilitzen diferents tècniques que s'aprofiten dels recursos presents en el vídeo com ara àudio, imatges facials o text.

1.1. Declaració d'intencions

L'objectiu d'aquest projecte és explorar tècniques basades en *deep learnig* i la implementació d'un sistema de verificació per tal de millorar el rendiment de l'anotació no supervisada en seqüencies de vídeo.

El treball es centrarà en la identificació de persones mitjançant el reconeixement de noms escrits i la identificació facial ja que són dels recursos més rellevants que s'utilitzen en les anotacions multimodals.

Aquest treball es desenvolupa a l'entorn del CAMOMILLE project (Collaborative Annotation of multi-MOdal, multI-Lingual and multi-mEdia documents) proposat per l' European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-NET.

El projecte neix com a la continuació lògica de la submissió de l'any passat per part de la UPC, amb l'objectiu millorar-ne els resultats mitjançant l'ús de noves tècniques d'anotació.

Per dur a terme aquesta tasca s'implementarà un sistema complert d'anotació automàtica i es desenvoluparan nous components per tal d'incorporar l'ús de tècniques de *deep learning*. Finalment es farà una comparativa dels diferents mètodes proposats per seleccionar-ne els millors i acabar comparant-ne el rendiment amb la submissió de l'any passat, UPC System for the 2015 MediaEval Multimodal Person Discovery in Broadcast TV task[1].

Les principals contribucions que s'han realitzat amb aquest projecte són les següents:

- Desenvolupament de un sistema de verificació per l'anotació de vídeos
- Desenvolupament de un sistema de clustering per l'anotació de vídeos
- Desenvolupament d'un sistema extractor de característiques facials mitjançant CNN
- Desenvolupament d'un programa d'anotació manual

1.2. Requeriments i especificacions

Requeriments del projecte:

- Desenvolupar un sistema d'extracció de característiques facials mitjançant CNN
- Desenvolupar un programa base per el sistema d'anotació automàtica
- Provar i desenvolupar els diferents mètodes de verificació
- Provar i desenvolupar un sistema basat en clustering

Especificacions del projecte:

- Els programes d'anotació automàtica i manual estan desenvolupats amb Python.
- El programa d'extracció de característiques amb CNN esta desenvolupat amb Matlab
- El programa d'anotació automàtica monomodal basat en vídeo ha d'assolir un rendiment superior a l'utilitzat en la submissió de l'any passat

1.3. Pla de Treball

Durant el desenvolupament del projecte s'ha intentat seguir la planificació marcada per el *project proposal plan* i el *critical review*. El pla de treball proposat té com a objectiu familiaritzar-se amb la temàtica del projecte i les eines principals per després centrar-se exclusivament en el desenvolupament.

En general, a excepció de l'addició del desenvolupament del software d'anotació manual i alguns petits ajustaments en els terminis de les tasques, s'ha seguit la planificació establerta.

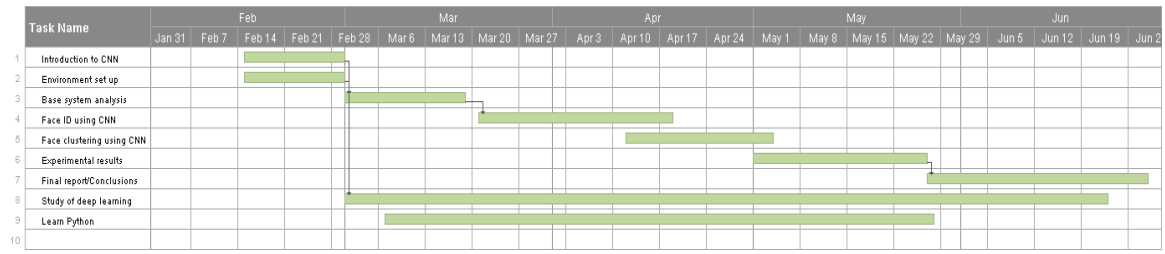


Figura 1: Diagrama de Gantt proposat en el project proposal plan

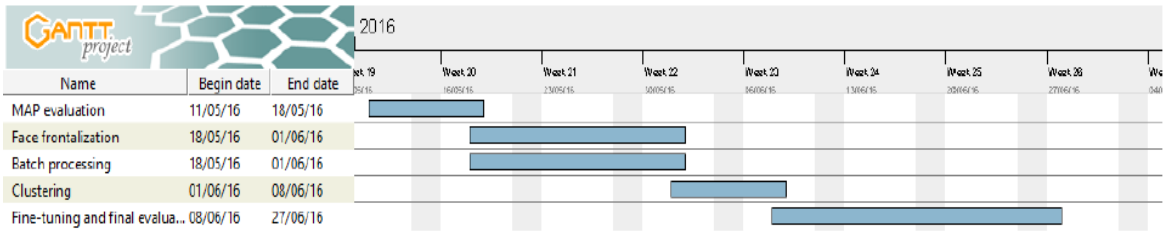


Figura 2: Diagrama de Gantt proposat en el critical review

2. Estat de l'art:

2.1. Xarxes neuronals convolucionals (CNN'S)

Les xarxes neuronals convolucionals són un sistema de *machine learning* amb una certa semblança estructural al funcionament de còrtex visual humà. Aquest tipus de xarxa són una variant de les xarxes neurals multicapa on la seva arquitectura està optimitzada per treballar amb entrades 2D, de manera que són molt més efectives per tasques com la visió artificial.

Aquestes xarxes tenen l'objectiu de extreure informació d'alt nivell de les imatges d'entrada per finalment classificar-les o extreure'n els vectors de característiques. Per tal de realitzar tot aquest procés, les diferents capes que conformen al xarxa neuronal s'encarreguen de processar la imatge d'entrada de manera que la sortida de l'ultima capa s'obtingui el resultat. Les capes estan organitzades de manera que cada una d'elles processa la sortida de l'anterior excepte la primera capa que processa directament la imatge i l'ultima capa on la sortida ja és directament el resultat.

Actualment podem trobar una gran varietat de capes segons el tipus d'arquitectura utilitzada. Tot i això normalment acostuma a haver 4 tipus de capes bàsiques presents en la majoria d'elles:

- ***Convolutional Layer:*** És la capa bàsica que constitueix a xarxa. Els paràmetres d'aquesta capa són un conjunt de filtres que realitzen una convolució sobre la informació d'entrada. De manera que si tenim k filtres de dimensions $n_x n_x r$ que operen sobre unes dades d'entrada $m_x m_x r$ obtindrem una sortida amb dimensions $(m-n+1)_x(m-n+1)_x k$. Un dels paràmetres d'aquest tipus de capes es el pas de convolució (*stride*), que és equivalent aplicar un delmat a la sortida de la convolució.

- ***Pooling Layer:*** L'objectiu d'aquesta capa es reduir la mida espacial de l'entrada per tal de disminuir el nombre de paràmetres i de càlculs. Actualment hi ha diferents criteris per realitzar aquesta operació dels quals el *max pooling* és el més utilitzat. Aquest consisteix en transformar regions 2x2 o 4x4 en una sola component d'igual valor a l'element màxim de l'entrada.

- ***Fully Connected Layer:*** Mitjançant una convolució 1x1 aquest tipus de capes s'encarreguen de generar una sortida mono-dimensional. Normalment es posiciona en les ultimes posicions i s'encarreguen de realitzar classificació o raonaments d'alt nivell. Aquest tipus de capes no deixen de ser un cas particular de capes convolucions.

- ***ReLU Layer (non-linearity):*** La seva funció es introduir no-linearitat en la xarxa, per exemple mitjançant la funció $f(x)=\max(0,x)$ que resulta en un llindar d'activació a 0.

Per tal de determinar els paràmetres dels filtres de les capes convolucions s'ha de realitzar un entrenament de la xarxa sobre un conjunt de dades etiquetades destinades a l'entrenament. L'entrenament és realitza mitjançant el procés de *back propagation* on per a cada imatge d'entrada es computa l'error que s'ha produït en la classificació i s'actualitzen el paràmetres dels filtres per tal de minimitzar-lo mitjançant l'algorisme de *gradient descent*.

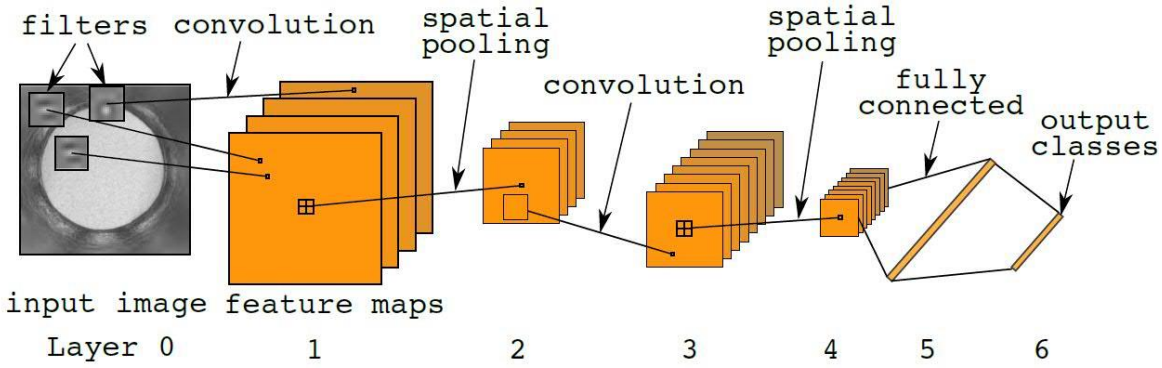


Figura 3: Exemple d'estructura de xarxa neuronal convolucional, on és pot observar la connexió típica entre capes i els diferents elements que la conformen.

2.2. Deep Face VGG-NET

VGG-NET es una arquitectura de xarxa neuronal convolucional proposada per Karen Simonyan & Andrew Zisserman del Visual Geometry Group d'Oxford. Aquesta xarxa va aconseguir la primera posició en la ImageNet Challenge 2014 [2].

Les principals característiques d'aquesta xarxa són les següents:

- Les imatges d'entrada es pre-processen mitjançant la subtracció de la mitja RGB computada en el conjunt d'entrenament per a cada píxel.
- Els filtres utilitzats en les capes convolucionals presenten dimensions bastant reduïdes 3x3 que (les mínimes dimensions que permeten captura la noció de posició)
- El pas de convolució és manté fixe a 1 i totes les capes de convolució estan seguides per un ReLU.
- En les capes de convolució és realitza *padding* per tal de preservar la resolució de sortida.
- S'utilitza *max-pooling* sobre regions 2×2 amb un *stride* de 2, no totes les capes convolucionals estan precedides per *pooling*.
- La xarxa finalitza amb tres *fully conected layers* (FC): les dues primeres tenen 4096 connexions mentre que la tercera realitza la classificació *softmax* amb 1000.

La xarxa VGG té diferents variants, la utilitzada en aquest projecte es la variant A. La seva estructura es mostra a continuació.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Taula 1: Estructura de la configuració A de la xarxa VGG. Taula extreta de [3]

Aquesta arquitectura s'ha entrenat amb èxit per tal de realitzar tasques de verificació i classificació facial. Aconseguint una exactitud del 98.95% i 97.3% en les bases de dades Labeled Wild Faces(LWF) i Youtube Faces Dataset respectivament [3].

L'entrenament es va realitzar mitjançant una base de dades composta per 2.6 milions d'imatges provinents de 2622 identitats amb l'objectiu de realitzar classificació i verificació facial. De manera que es poden obtenir els resultats de classificació o els vectors de característiques a partir de qualsevol imatge facial d'entrada.

2.3. Identificació no supervisada de persones en vídeos

La identificació no supervisada de persones en seqüencies de vídeos és una tasca amb un llarg recorregut. Al 1999 és van realitzar les primeres propostes per afrontar el problema [4],[5].

L'objectiu de la identificació no supervisada consisteix en anotar els moments d'aparició per pantalla de les diferents persones que apareixen en una seqüencia de vídeo. Per tant, per una banda és requereix esbrinar els noms de les persones mitjançant la informació present en el vídeo, i per altre banda, identificar tot els moments en que una determinada persona hi apareix.

La majoria dels sistemes actuals són multimodals, és a dir que utilitzen diferents fonts de informació per realitzar aquestes tasques. Actualment les 3 fonts d'informació més comunes són els noms escrits que apareixen per pantalla, l'àudio corresponent als segments de parla i les característiques facials de les persones. Cada una d'elles s'utilitza per realitzar una tasca concreta.

- Noms escrits: Aquesta font d'informació s'utilitza per associar els noms que apareixen en el vídeo a les persones que apareixen juntament amb ells. Obtenint així unes primeres hipòtesis d'identificació que serviran com a base per identificar la resta de persones.
- Característiques facials: Aquests descriptors s'extreuen a partir de les cares detectades en els fotogrames del vídeo i serveixen com a identificador biomètric per a cada identitat. S'utilitzen per segmentar els intervals d'aparició d'una mateixa persona [6].
- Fragments de parla: L'àudio present en un vídeo pot aportar 2 tipus d'informació. Per una banda permet realitzar una primera identificació de les persones a partir dels noms esmentats en la parla i per altre banda també permet realitzar una dialització dels locutors utilitzant la veu com a model biomètric per identificar les persones de manera semblant a l'ús de les cares. [7],[8].

Per aprofitar els recursos presents en aquestes fonts d'informació s'utilitzen tècniques molt diverses i que en general, han seguit una notable evolució per tal d'augmentar-ne la fiabilitat.

Una altre part molt important en l'anotació no supervisada és la metodologia utilitzada per realitzar les segmentacions, fusió d'informació i propagació de noms. Normalment,

s'utilitzen tècniques de clustering tan per la diarització de parla com per la segmentació facial tot i que també es poden utilitzar altres metodologies com la utilització de grafs [9].

2.4. Gaussian Naive Bayes

En *machine learning,* els classificadors *Naïve Bayes* corresponen a una família de classificadors probabilístics que és basen en l'aplicació del teorema de Bayes amb una forta assumpció d'independència entre característiques [10].

Donada una variable y i un vector x_1, ...,x_n associat, el teorema de Bayes exposa la següent relació:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Mitjançant l'assumpció de que els components de x són independents, el criteri MAP (Maximum a Posterior) permet classificar-lo de la següent forma:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

En el cas gaussià, s'assumeix una distribució de probabilitat Gaussiana en les característiques.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

De manera que mitjançant un conjunt d'entrenament amb les etiquetes de cada classe podrem calcular P(y) i els paràmetres de la distribució gaussiana de cada classe.

D'aquesta manera obtenim un classificador basat en el criteri MAP on la decisió es realitzarà mitjançant l'avaluació del *log likelihood ratio.*

2.5. Joint Bayesian

El *joint Bayes* és una proposta de classificador binari per tal de determinar si dues cares pertanyen al mateixa persona o no.

Aquest mètode ha obtingut uns resultats del 92.4% d'exactitud en la *Challenging Labeled Face in Wild (LFW) data set*. El que suposa una reducció de un 10% comparat amb els mètodes convencionals [11].

A diferencia de altres mètodes clàssics que treballen sobre la diferencia entre dues cares com es mostra en la figura 4, el *joint bayesian* intenta modelar la distribució conjunta entre dues cares (x1,x2) a avaluar per tal d'evitar reduir-ne la separabilitat.

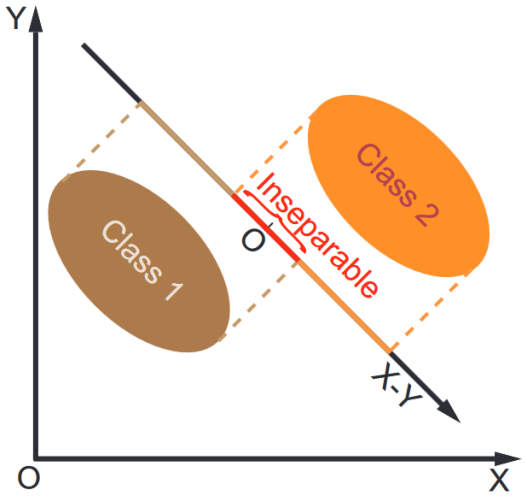


Figura 4: Il·lustració de la pèrdua de separabilitat entre 2 objectes a causa de la reducció de la dimensionalitat mitjançant mètodes convencionals. Imatge extreta de [11]

Per aconseguir les distribucions conjuntes de (x_1,x_2) es defineixen a priori les representacions facials com a la suma de dues variables independents gaussianes

$$x = \mu + \varepsilon$$

on x correspon a la cara observada, μ s'encarreguen de modelar la part intrínseca de la identitat i ε la variació intra-personal (expressions, il·luminació, etc...). μ i ε segueixen dues distribucions gaussianes $N(0, S_\mu)$ i $N(0, S_\varepsilon)$ on S_μ i S_ε són les dues matrius de covariàncies desconegudes.

Basant-nos en aquestes assumpcions, mitjançant algoritmes de maximització de l'esperança es poden aprendre efectivament els models paramètrics de les dues variables, fent possible la obtenció de les distribucions conjuntes.

Un cop obtingudes les distribucions, es possible establir la relació de logaritme de les probabilitats amb la següent expressió. On el numerador correspon a la probabilitat de que x_1 i x_2 pertanyin a la mateixa identitat i el denominador a que x_1 i x_2 pertanyin a identitats diferents.

$$r(x_1, x_2) = \log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)} = x_1^T A x_1 + x_2^T A x_2 - 2x_1^T G x_2$$

Per tant A i G seran les matrius a determinar en la fase d'entrenament per tal de realitzar la verificació.

Finalment, establint un llindar sobre la relació anterior podrem decidir si un conjunt de 2 cares pertany a la mateixa identitat o no.

2.6. Clustering Jeràrquic

La clusterització de dades o clustering és una tècnica molt utilitzada en anàlisis de dades i que té com a objectiu agrupar una col·lecció d'objectes de tal manera que els elements d'un mateix grup (anomenat clústers) siguin mes semblants entre ells que no pas amb els dels altres grups.

En concret, el clustering jeràrquic correspon a la família d'algoritmes de clustering on es construeixen clústers a partir de la unió o separació successiva de les dades. Un clúster jeràrquic es representa en forma d'arbre on les arrels de l'arbre corresponen a un únic

clúster que agrupa tots els elements i les fulles s'associen a clústers amb un únic element.

Per tal de generar l'arbre complert, normalment es fusionen successivament els elements més similars (utilitzant el mètode i la mètrica seleccionada) per tal de generar nous clústers. De manera que s'obté un arbre complert amb totes les agrupacions com a resultat final.

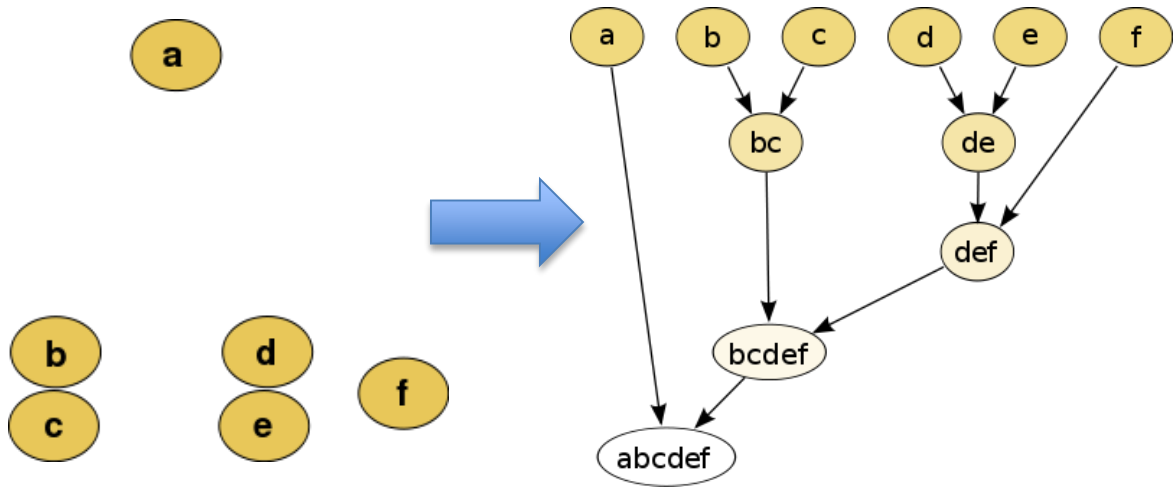


Figura 5: Il·lustració que mostra com a partir de un conjunt d'elements inicials a,b,c,d,e,f és poden realitzar agrupacions entre ells en funció de la distancia utilitzant la tècnica de clustering jeràrquic. Imatge extreta de [12]

Llavors mitjançant la poda d'aquest arbre s'aconsegueix agrupar les dades d'entrada en les agrupacions desitjades.

Normalment en aplicacions no supervisades on el nombre de clústers o la distancia màxima entre elements és un paràmetre desconegut s'utilitzen diferents tècniques d'anàlisi per tal de reconèixer el nombre òptim de clústers. Com ara el mètode de silueta [13] o el criteri de Calinski-Harabasz [14].

Calinski Harabasz és basa en la maximització del criteri de relacions de variància (VRC) per tal de trobar en número de clústers òptims. És busquen aquells clústers que estan ben definits i que per tant presenten una variància entre clústers alta i una variància dins el mateix clúster baixa.

El mètode de silueta en canvi mesura com de similar són els punts dins d'un mateix clúster en comparació als punts residents en altres clústers.

3. Desenvolupament del projecte/Metodologia:

El projecte consta de dues aplicacions que s'han desenvolupat durant el curs. Per una banda el programa principal, que té l'objectiu de realitzar la anotació no supervisada de seqüencies de vídeo, i per altre banda, una aplicació auxiliar que ha sorgit davant la necessitat de disposar de seqüencies de vídeo anotades manualment per tal de entrenar i avaluar el rendiment del programa principal.

3.1. Sistema d'anotació no supervisada

El programa principal té una estructura modular. Esta compost per diferents components que s'encarreguen de realitzar les diferents etapes de tot el procés d'anotació. D'aquesta manera s'augmenta la flexibilitat i la intercanviabilitat entre mòduls.

El programa és pot separar en 3 grans blocs que consten de pre-procesament, extracció de característiques i identificació. Aquests s'interconnecten de la següent manera:

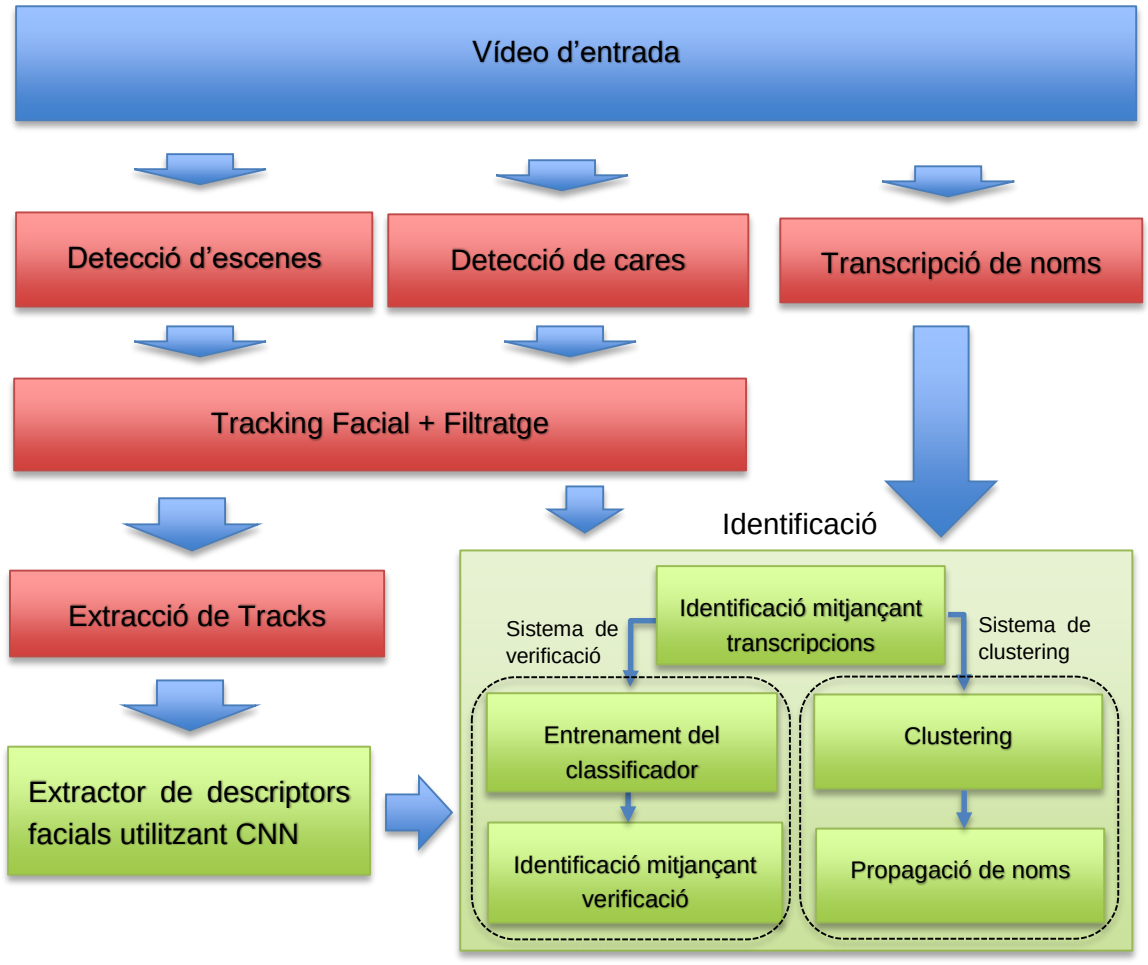


Figura 6: Diagrama de cconnexions que mostra l'estructura del sistema. En verd les etapes que s'han desenvolupat íntegrament durant el projecte.

3.1.1. Pre-procesament

El pre-procesament té com objectiu extreure tota la informació base per els sistemes posteriors a partir dels fitxers de vídeo. Per realitzar aquesta tasca s'utilitzen diferents eines i scripts que ja s'utilitzaven en el projecte anterior.

A partir del vídeo d'entrada, el primer pas a realitzar és la detecció d'escenes mitjançant un programa anomenat Shot Detector Maseter. Aquest software analitza la quantitat de moviment entre fotogrames i mitjançant un llindar definit com a paràmetre d'entrada decideix en quins instants del vídeo es produeix els canvis d'escena.

El següent pas consisteix en la detecció de cares mitjançant un script de Python. Aquest script, utilitza les llibreries open CV per tal de detectar les cares que apareixen durant el vídeo mitjançant un detector Viola–Jones entrenat específicament per aquesta tasca. Amb aquest script s'aconsegueix un índex de totes les cares detectades en cada fotograma juntament amb les seves dimensions i posició.

A partir d'aquí s'utilitza un altre script que s'encarrega de definir els intervals temporals en que apareix una mateixa persona. Aquests intervals temporals se'ls assigna una identificació única i se'ls anomena tracks. El criteri basic per decidir si una seqüencia consecutives de deteccions facials correspon o no a un track és basa en l'estudi del flux òptic. Ja que duran la creació de tracks es produeixen molts falsos positius, és realitza un filtratge posterior per tal de eliminar tracks amb alta probabilitat de ser falsos positius com ara els de longituds massa curtes.

Finalment un nou script s'encarrega d'extreure les cares detectades per a cada track per tal de utilitzar-les durant la verificació facial. D'aquesta manera, cada track tindrà associades un conjunt de cares corresponents a la identitat que representa. Aquest mateix script també disposa de la opció de frontalitzar-les. Un procés on es sintetitza una visó frontal de la cara a partir de les imatges d'entrada [15].

Un altre part important consisteix en realitzar les transcripcions dels noms que apareixen durant el vídeo. Per aconseguir-ho primer es realitza una detecció dels fragments de text, llavors és transcriuen mitjançant tècniques OCR i finalment s'hi busquen aparicions de noms propis a partir de un diccionari. Els noms detectats serviran per identificar aquells tracks que han aparegut en el mateix interval temporal, i que llavors serviran per identificar la resta.

Figura 7: Exemple de l'aparició d'un nom escrit que permetrà identificar a la persona que esta parlant

Aquest conjunt d'elements de pre-pocessament han estat proporcionats com a sistema base per els organitzadors MediaEval 2015, a excepció del sistema de frontalització desenvolupat per [15].

3.1.2. Extracció de Característiques

En aquesta nova etapa, l'objectiu es extreure vectors de característiques de les imatges facials de cada track, per facilitar-ne la seva identificació en fases posteriors. Aquesta tasca és realitza mitjançant una xarxa de convolucions neuronals implementada sobre MatConvNet.

MatConvNet és una plataforma pensada per treballar amb CNN que presenta una gran flexibilitat a l'hora de treballar gràcies a un sistema d'estructura per capes i que permet treballar utilitzant tant la GPU com la CPU. Aquesta plataforma disposa de totes les operacions i estructures de les xarxes neurals implementades, i permet carregar amb facilitat diferents arquitectures de xarxes.

Degut al poc temps del projecte i la complexitat per entrenar una xarxa d'aquestes característiques s'ha decidit utilitzar una xarxa pre-entrenada per tal de realitzar aquesta tasca. En concret l'arquitectura VGG entrenada específicament per a classificació facial. Cal destacar que tot i que l'entrenament de la xarxa s'hagi realitzat mitjançant imatges facials de persones no presents en les seqüencies de vídeo a analitzar, els vectors de característiques que s'obtenen són prou representatius i discriminatius.

Al disposar d'una xarxa amb tots els pesos inicialitzats, només s'ha de carregar l'estructura en memòria i acte seguit ja es poden començar a extreure els vectors de característiques posant com a entrada la imatge desitjada i accedint a la sortida de la penúltima capa.

Aquest procediment ens permet obtenir els descriptors, ja que al tractar-se de una xarxa entrenada amb l'objectiu de classificació, la xarxa transforma la imatge original en noves estructures de cada cop més alt nivell fins a arribar a l'ultima capa de classificació. És per això que ens interessa accedir a aquesta penúltima capa que és on hi ha les descripcions de més alt nivell de la imatge.

Per tal de integrat aquesta xarxa en el sistema complert, s'ha desenvolupat un script Matlab que extreu tots els descriptors facials donat un conjunt de tracks.

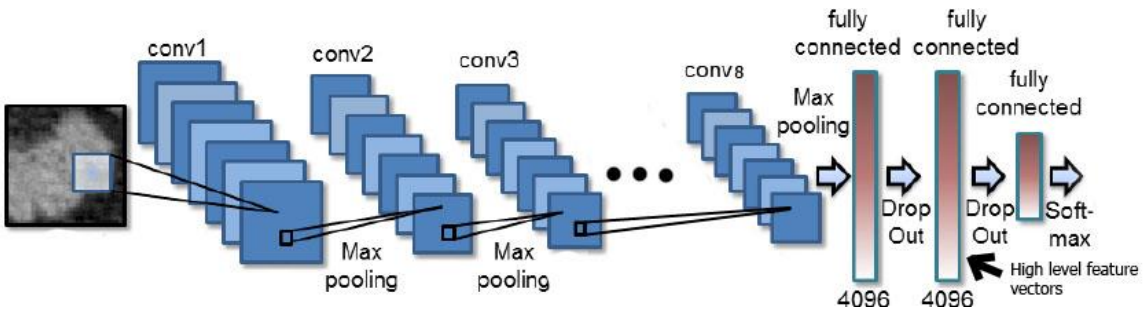


Figura 8: Imatge on es mostra la estructura de la xarxa utilitzada. Els vectors de característiques s'obtenen a la sortida de la penúltima capa. Imatge modificada a partir de [16]

3.1.3. Verificador

Aquesta última etapa es tracta d'un programa realitzat amb Python que té l'objectiu de generar els resultats d'anotació mitjançant verificació facial dels tracks extrets i les transcripcions dels noms.

El mecanisme general de funcionament es el següent:

1) És carreguen els vectors de característiques de cada track i es crea un objecte de tipus track que conté tota la informació necessària (frame inical, frame final, track ID, etc...)
2) Es busquen solapaments entre l'aparició dels noms transcrits i l'aparició dels tracks per tal d'identificar els tracks que apareixen juntament amb un nom. Aquests tracks identificats s'utilitzaran per entrenar el classificador i identificar la resta de tracks. A partir d'aquí els anomenarem TOK.
3) Cada track no identificat es compara un per un amb tots els TOK, passant-los per el classificador per mirar si els dos tracks corresponen a la mateixa persona o no. En cas de que es tractin de la mateixa persona, s'identificà el nou track amb el mateix nom que el track procedent de TOK.
4) Finalment, es comparen les anotacions fetes per el programa amb les anotacions manuals per tal d'avaluar el rendiment del sistema.

Malgrat que el funcionament general es sempre el mateix, es possible realitzar les diferents etapes de varies maneres diferents. Per aquesta raó, s'ha decidit provar amb diferents mètodes per a cada cas per tal de trobar la configuració optima.

3.1.3.1. Verificació facial

Per decidir si dos imatges facials corresponen o no a la mateixa identitat s'utilitza un classificador. Aquest decideix si els 2 vectors de característiques d'entrada corresponen a la mateixa persona o no.

Per el sistema de verificació s'han provat 2 classificadors diferents: El *Gaussian Naive Bayes* i el Joint Bayesian.

***Gaussian Naive Bayes*:** Aquest classificador es capaç de decidir si el resultat de la distancia euclidiana entre dos vectors correspon a una distancia entre elements d'una mateixa persona (distancia intra) o a una distancia entre elements de persones diferents (distancia inter). Per tant es tracta d'un classificador binari que utilitza una característica unidimensional, la distancia euclidiana. Per tant s'haurà d'establir un llindar sobre la distancia per determinar si un parell de cares corresponent a la mateixa identitat. El llindar es decidirà en funció de les probabilitats a priori i les funcions de probabilitat calculades a partir de la seqüencia d'entrenament.

Abans de començar amb la identificació dels tracks sense text solapat, s'utilitzen els TOK per tal d'entrenar aquest classificador. Es calculen totes les distancies entra cares pertanyents al mateix track i totes les distancies entre cares entre tracks diferents, aquestes s'etiqueten respectivament amb la finalitat de obtenir una seqüencia d'entrenament.

Seguint aquesta metodologia, s'obté un nombre molt superior de distancies inter que de distancies intra, ja que la quantitat de distancies per classe segueixen les següents expressions:

$$\#inter = Nm^2 \qquad \#intra = N\binom{m}{2}$$

Per tant, un cop entrenat el classificador, es canvien les probabilitats a priori per unes de més properes a les esperades a l'hora de verificar els tracks. En totes les implementacions que s'utilitza el classificador Bayesià s'han utilitzat les probabilitats a priori calculades a partir de una seqüencia d'entrenament de 3 vídeos anotats manualment. La probabilitat de que la distancia correspongui a intra s'ha calculat mitjançant la formula escrita a continuació per a cada vídeo i llavors s'ha calculat la mitja entre els 3 vídeos d'entrenament:

$$P(intra) = \frac{1}{\#tracks\_ok} \cdot \frac{\#track\_identificats}{\#tracks\_ok}$$

***Joint Bayesian**:* Aquest classificador també té l'objectiu de decidir si dos cares corresponen a la mateixa persona o no.

En una primera etapa, es va intentar entrenar el classificador de la mateixa manera que el Naive Bayesian, utilitzant les cares de els persones amb text solapat com a entrenament (TOK). El problema es que es disposa únicament de unes 30 identitats amb unes 30 cares per identitat i al tractar-se de seqüències de vídeo, la variabilitat és baixa (les imatges facials dins un track son molt similars). En conseqüència l'entrenament resulta insuficient donat les elevades dimensions dels vectors de característiques.

Per tant es va decidir que s'entrenaria el classificador utilitzant una base de dades externa de dimensions suficients que s'utilitzaria per igual en tots els vídeos a anotar.

Seguint aquest procediment s'obté tota la part de les distribucions conjuntes mitjançant les covariàncies de la base de dades externa i s'utilitzen els vídeos de la seqüencia d'entrenament per determinar el llindar de discissió.

També s'ha provat una versió que utilitza anàlisis de components principals (PCA) amb l'objectiu de reduir la dimensionalitat dels vectors de característiques i així millorar la fase d'entrenament.

3.1.3.2. Verificació de tracks

Donat que cada track conté una seqüencia d'imatges facials associades, existeixen diferents possibilitats a l'hora de realitzar la verificació a nivell de track. Per tal decidir el millor mètode s'han provat i avaluat diferents alternatives.

En tots els casos, per a cada track no identificat que es compara amb tots els TOK, s'acaba identificant amb el track que ha obtingut la puntuació més alta en cas de que s'hagi superat el llindar d'identificació.

- **Tots contra tots:** Aquest mètode consisteix en computar la distancia per totes les parelles de vectors facials entre els dos tracks a avaluar. Llavors hi ha 2 variacions possibles de l'algorisme: *hard decision* i *soft decision*.

  En la opció de *hard decision*, és passen totes les distancies calculades pel classificador, s'obté la classe per a cada una d'elles, i finalment s'estableix un segon llindar per decidir quina relació hi ha d'haver entre els resultats de cada classe per decidir si els 2 tracks corresponen a la mateixa identitat o no.

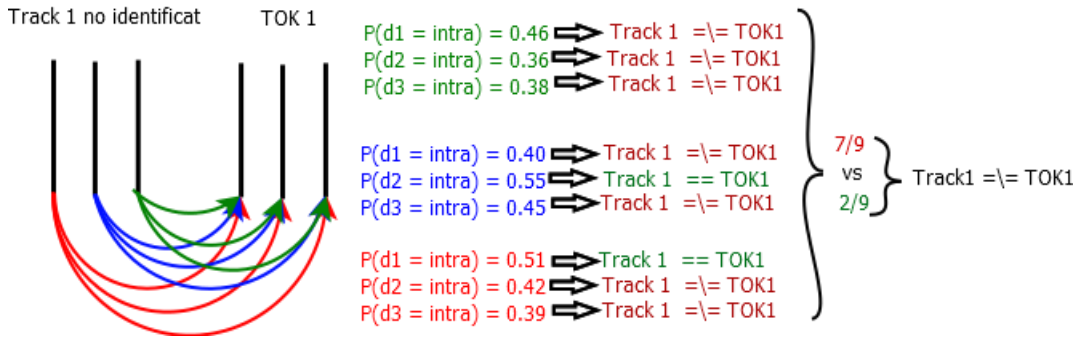


Figura 9: Esquema que mostra el criteri de verificació entre 2 tracks utilitzant el mètode de tots contra tots mitjançant hard descision

En la opció de *soft descision*, en comptes de predir la classe per a cada distancia, el que es fa es obtenir les probabilitats de que cada distancia sigui de una classe concreta. Llavors es fa la mitja de totes aquestes probabilitats per obtenir la probabilitat de que els dos tracks pertanyin a la mateixa persona o no.

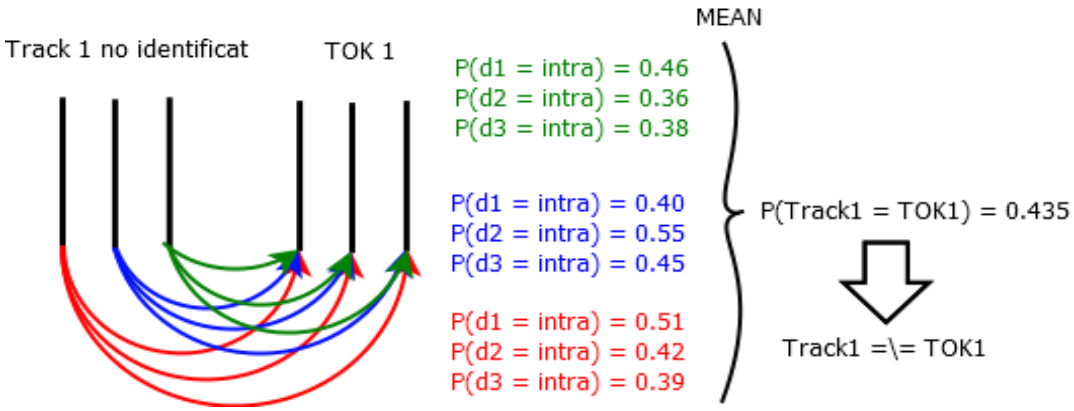


Figura 10: Esquema que mostra el criteri de verificació entre 2 tracks utilitzant el mètode de tots contra tots mitjançant soft descision

En totes dos opcions, es pot reduir el nombre màxim de vectors per track a analitzar durant la verificació per tal d'augmentar-ne la velocitat de execució.

- **Vector més representatiu/distancia mínima:** Aquest mètode consisteix a seleccionar un vector de característiques que representi cada track per realitzar la verificació. Per tant s'escull el vector que minimitzi la suma de diferencies amb la resta de tracks.

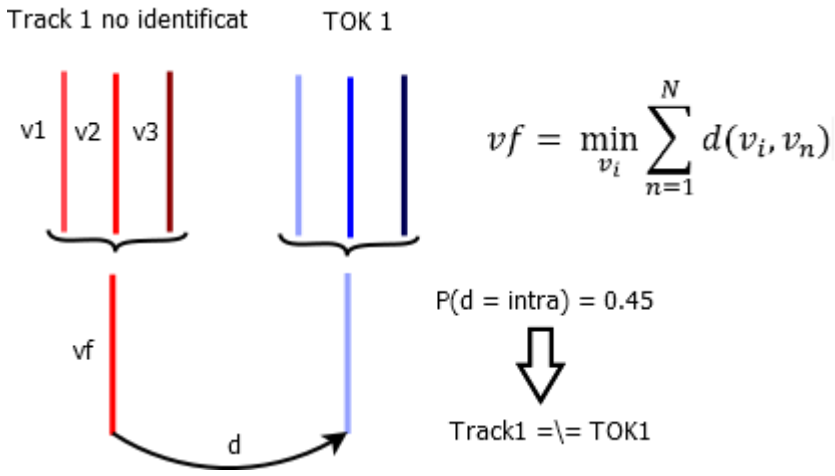


Figura 11: Esquema que mostra el criteri de verificació entre 2 tracks utilitzant el mètode del vector més representatiu/distancia mínima.

- **Mitjana de vectors de característiques:** Amb aquest mètode es genera un nou vector de característiques per tal de representar el track. Aquest nou vector s'obté a partir de realitzar la mitjana de tots els vectors associats.

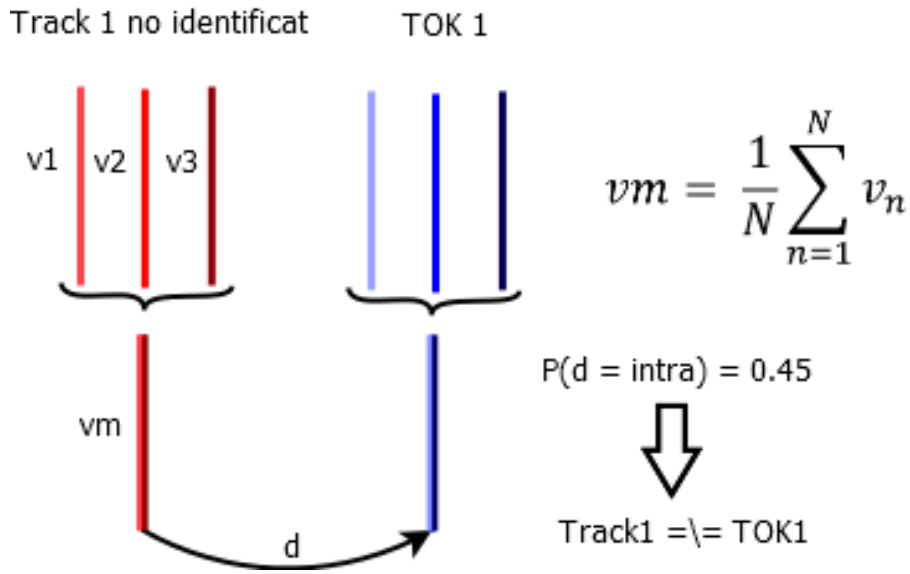


Figura 12: Esquema que mostra el criteri de verificació entre 2 tracks utilitzant el mètode del mitja vectorial.

Quan un vector no identificat es compara amb tots els TOK se l'hi assigna un valor de confiança corresponent a la decisió que s'ha pres a l'hora de classificar-lo i que més endavant s'utilitzarà en la mètrica d'avaluació. Cada mètode té un criteri diferent per calcular aquesta confiança, però en general és proporcional a les probabilitats de que el track analitzat correspongui a la mateixa identitat que el track de referencia..

3.1.3.3. Sistema d'identificació mitjançant transcripcions

Un dels problemes principals durant la identificació mitjançant les transcripcions de text succeeix quan es produeix un solapament temporal entre varis trakcs. És a dir, que apareix més de una persona durant l'interval d'aparició del text. Aquests solapaments

múltiples és poden agrupar en 3 casos diferents en els quals s'han plantejat diferents mètodes per afrontar el problema.

- **Solapament seqüencial:** En aquest cas, durant l'interval d'aparició d'un text, apareixen múltiples tracks però no es solapen entre ells ja que apareixen de manera seqüencial. En aquest cas s'observen 2 possibles escenaris; que els diferents tracks pertanyin a una mateixa identitat (ja sigui per problemes de tracking o oclusió momentània) o que els tracks pertanyin a persones diferents. En el primer escenari es possible assignar el nom a els tracks ja que es tracta de la mateixa persona mentre que en el segon cas, sense disposar de la informació d'àudio, no es possible.

  Per tant l'objectiu d'aquest escenari consisteix en detectar si es tracta del primer o segon cas. Per dur a terme aquesta tasca, és podria utilitzar la diferencia de posició relativa entre la ubicació de les cares de cada track i també avaluar-ne la semblança mitjançant la comparació dels vectors de característiques.

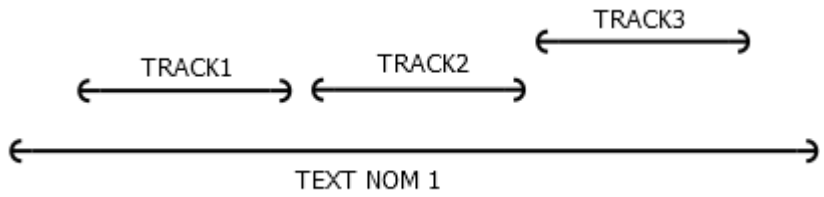


Figura 13: Il·lustració que representa la distribució temporal dels tracks i aparició del text en un cas de solapament seqüencial. Cada eix horitzontal representa una mateixa identitat.

- **Solapament simultàni:** En aquest cas, durant l'interval d'aparicio del text, més de una persona apareix simultàniament. Ja que el nostre sistema no disposa de detecció d'expressions facials per detectar quin track està parlant no es possible identificar el track.

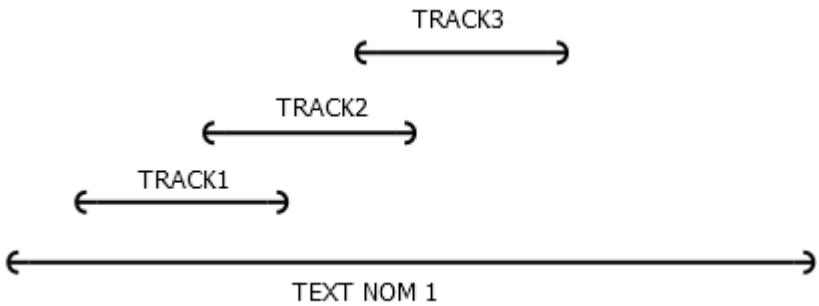


Figura 14: Representació de la distribució entre texts I tracks en el cas de solapament simultani.

- **Solapament mixta:** En aquest cas apareixen tan tracks solapats com tracks no solapats. Mitjançant els procediments esmentats anteriorment, es possible identificar tracks no solapats com a un mateix track i convertir el cas de solapament mixta en solapament simultani.

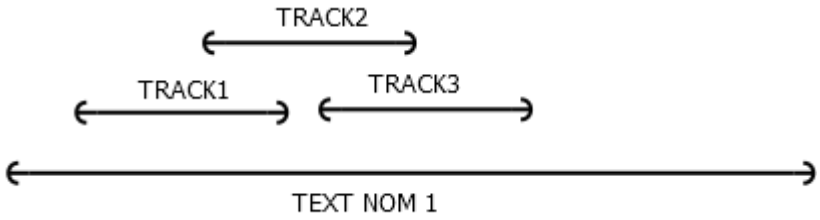


Figura 15: Representació de la distribució entre texts I tracks en el cas de solapament mixte

En tots 3 casos, sempre es possible acabar assignat el text a un track determinat utilitzant criteris probabilístics basats en la duració del track o ordre d'aparició.

Tot i així, després de provar amb diferents configuracions s'ha decidit no contemplar cap d'aquests casos i assignar el nom de la transcripció a tots els tracks solapats dins de l'interval.

La raó d'aquesta decisió es la següent: Normalment, a les persones que se'ls assigna un nom erroni mitjançant aquests procediments corresponen a espectadors o persones poc rellevants que no acostumen a aparèixer en futures trames del vídeo. Per tant, l'error que es coment amb aquestes assignacions incorrectes difícilment es propagarà en les següents etapes de verificació. Cometre aquests petits errors és menys crític que no pas deixar un personatge rellevant sense assignació de nom, ja que aquests acostumen a aparèixer diverses vegades de manera que no es podrien identificar la resta dels seus tracks i augmentaria la possibilitat de que aquests s'identifiquessin erròniament.

3.1.4. Anotació mitjançant clustering jerarquic

Aquest sistema d'anotació és un alternativa a la identificació mitjançant verificació i també té com a objectiu identificar una seria de tracks a partir de les cares extretes de cada track i les transcripcions dels noms. Aquest sistema però, es basa en un concepte totalment diferent, el clustering. L'objectiu és aconseguir agrupar tots els tracks d'una mateixa identitat en un sol clúster per facilitar la propagació de noms en cas de que algun dels tracks hagin estat identificat per les transcripcions de text.

En aquest cas s'ha decidit utilitzar un clustering jeràrquic ja que el número de clústers a realitzar dependrà del vídeo. Per definir els diferents clústers s'ha utilitzat la distancia euclidiana com a mètrica de distancia juntament amb el mètode de Ward. Aquest últim realitza les agrupacions entre elements mitjançant la minimització de la suma de les diferencies al quadrat. Un mètode que està basat el la minimització de variàncies.

De la mateixa manera que en el sistema de verificació, el primer pas consisteix en obtenir una llista amb tots els objecte track i realitzar les primeres identificacions a partir de les transcripcions de text.

Arribats a aquest punt, es realitza un clustering jeràrquic a partir de tot el conjunt de vectors de característiques i és defineixen els clústers amb la intenció de agrupar tots els vectors de cada identitat en un mateix clúster.

Finalitzat el clusterig, s'etiqueta a cada vector de característiques amb la id de clúster al que pertany.

Per tant, observant quina és la id del clúster predominant en el conjunt de vectors d'un mateixa track, podrem associar el track a un determinat clúster. A més a més, el

percentatge de vectors amb la id del clúster associat al track servirà com a valor de confiança.

L'últim pas consisteix en identificar els tracks que falten mitjançant la propagació de noms. En cas de que un clúster contingui a un track prèviament identificat, aquest identificarà a la resta amb el seu mateix nom, ja que en principi, tots els tracks dins un mateix clúster haurien de pertànyer a una mateixa identitat.

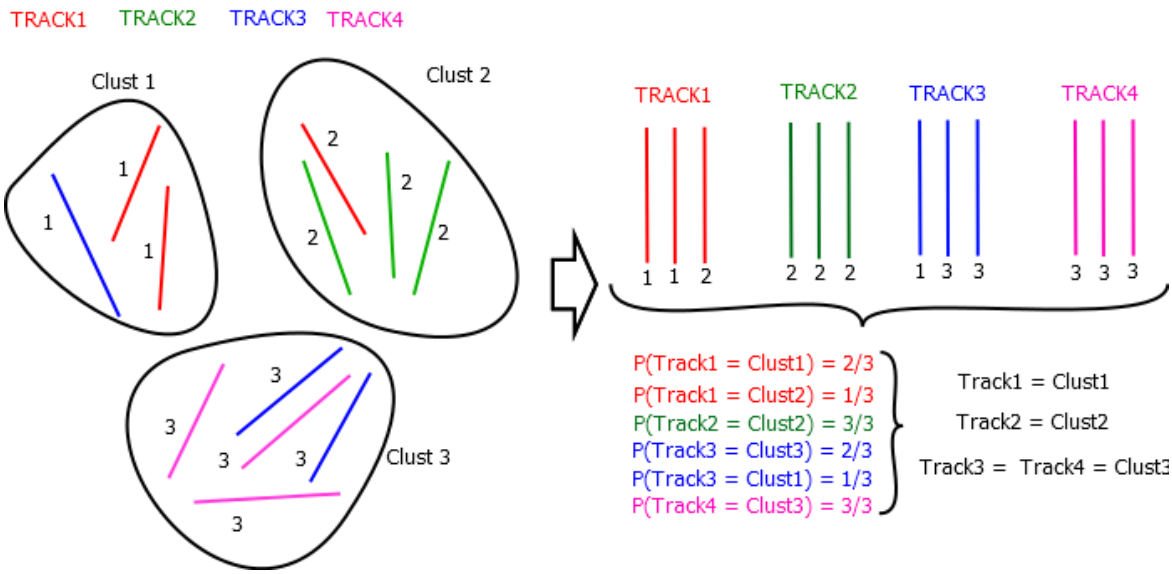


Figura 16 Il·lustració del procediment utilitzat en la identificació de tracks mitjançant el sistema de clustering.

En la Figura 16 es poden observar 3 casos diferents que il·lustren el procediment d'anotació mitjançant clustering:

- Track1: Dos dels seus vectors de característiques han anat a parar dins del clúster 1 mentre un altre ha anat a parar al clúster 2. Per tant el Track1 tindrà 2 vectors etiquetats amb la id1 i un altre amb la id3. Llavors, doncs que hi ha una relació favorable 2/3 respecte la id1 s'identificarà al track amb la id1.
- Track3 i 4: Amb el mateix procediment que el track1, tots dos tracks acaben obtenint la id3. Al compartir la mateixa id significa que els dos tracks pertanyen a la mateixa identitat i per tant es realitzarà la propagació de noms

3.2. Implementació del sistema 2015:

Per tal de poder comparar els sistemes desenvolupats en aquest projecte amb la submissió de l'any passat, s'ha decidit desenvolupar un altre software d'anotació mitjançant clustering jeràrquic, però aquest cop utilitzant els descriptors facials del sistema de l'any passat.

Inicialment aquest sistema és va desenvolupar amb un únic descriptor facial per track tal com es proposava en l'article de la submissió 2015. Tot i això al final s'ha decidit utilitzar diversos vectors per track per tal d'aconseguir un millor sistema de confiança. S'ha comprovat que les diferencies de rendiment en els 2 sistemes són pràcticament negligibles.

3.3. Avaluació del sistema d'anotació no supervisada:

Per tal de avaluar el rendiment final, s'ha decidit utilitzar un nou conjunt de vídeos anotats manualment com a seqüencies d'avaluació. Els paràmetres dels diferents mètodes s'han fixat a partir dels resultats de les corbes ROC buscant un equilibri entre maximitzar els TPR i minimitzar els FPR. Per tal de seleccionar el llindar òptim a partir del conjunt d'entrenament s'ha escollit el llindar que minimitza la expressió TPR-FPR. On TPR i FPR corresponen a el True Positive Rate i el False Positive Rate i s'obtenen a partir de les expressions següents:

$$TPR = \frac{tp}{tp+fn} \quad FPR = \frac{fp}{tp+tn}$$

La mètrica utilitzada per avaluar el rendiment final ha estat la mateixa que la que s'utilitzarà en la submissió del setembre, la mètrica MAP (*Mean Average Precision*).

La mètrica MAP és la mitjana aritmètica del conjunt de precisions mitjanes de totes les persones que apareixen en el vídeo. On per cada persona es realitza una consulta on s'obtenen les seves anotacions per odre de confiança. D'aquesta manera, les anotacions amb confiances altres són més rellevants en la puntuació final.

A continuació les formules per calcular la precisió mitjana i la MAP.

$$\mathrm{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k) \qquad \mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

On k és el rang en l'ordre dels elements retornats per la consulta, n és el nombre d'elements retornats, P(k) és la precisió en el conjunt d'elements que van des del primer fins a k, i Δ r (k) és canvi en el *recall* del conjunt d'elements que van des de k - 1 a k.

Tot i així s'ha de destacar que les anotacions manuals utilitzades per obtenir els resultats finals, s'han generat amb l'objectiu d'avaluar només l'etapa final d'anotació i no tot el sistema. Per tant les puntuacions obtingudes no es correspondran amb les que s'obtindrien al avaluar el sistema complert.

3.4. Software d'anotació manual

Per tal d'avaluar el rendiment d'anotació de cada sistema es requereix de taules de la veritat, on hi ha l'anotació correcta de tots els tracks. Aconseguir aquestes taules no es una feina senzilla ja que l'anotació manual es lenta. Així que per tal d'agilitzar aquest procés es va decidir crear un programa que assistís en aquesta tasca.

El programa mostra un per un el segment de vídeo associat a cada track juntament amb el requadre de la cara a on esta associat. L'usuari llavors pot decidir donar-li una id personal nova o associar-lo amb alguna ja existent. En cas de que aparegui un text per pantalla amb el nom de la persona que esta parlant, el programa suggerirà associar aquest nom a la identitat del track o permetrà a l'usari introduir un altre nom en cas d'errors. Al final de l'execució s'anota el nom de cada track.

El programa també disposa de diferents comandes que permeten algunes funcionalitats extra. Com ara repetir la reproducció d'un track en cas de que l'usuari no haguí prestat suficient atenció o tingui alguns dubtes, visualitzar un eix temporal amb una imatge de cada track per comprovar si alguna persona ha aparegut amb anterioritat o no i buscar quina es la seva id, o identificar un nou track amb la mateixa id del track anterior sense haver-la d'especificar.

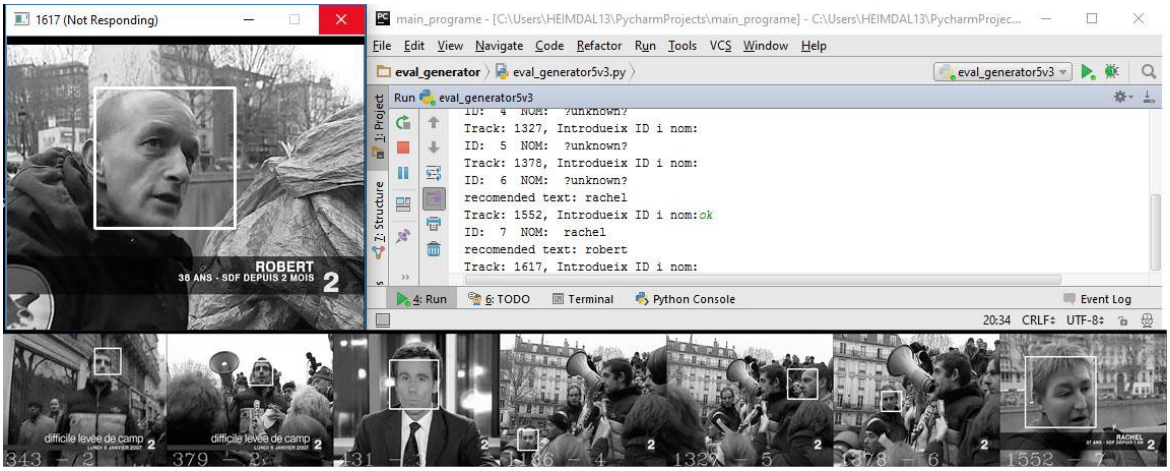


Figura 17: Imatge de la interfície del software d'anotació manual. A l'esquerra és mostra la reproducció del track actual. A la dreta la consola on s'introdueixen les comandes i finalment a la part inferior es mostra una imatge amb tots els tracks que s'han anotat anteriorment juntament amb la seva ID.

4. Resultats

Amb l'objectiu de maximitzar el rendiment del software d'anotació no supervisada s'han decidit comparar els resultats obtinguts amb diferents mètodes, classificadors i algorismes de verificació per tal de seleccionar-ne la configuració més adient.

La comparació dels diferents mètodes s'ha realitzant mitjançant la comparació de les corbes ROC dels resultats procedents del programà d'anotació. Utilitzant el TPR i el FPR obtingut a partir de la comparació de les anotacions no supervisades amb les de referencia obtingudes manualment.

Ja que l'objectiu d'aquesta part consisteix en avaluar les millores respecte el projecte anterior i no pas el rendiment general del sistema, les taules de la veritat no contenen les anotacions reals del vídeo sinó que només s'anoten aquelles persones les quals son possibles d'identificar mitjançant la informació provinent dels components de pre-procesament. D'aquesta manera, els resultats que s'obtenen depenen directament del rendiment de la part d'identificació no supervisada i no pas de les fases anteriors. Això implica que durant l'anotació manual dels vídeos s'ha seguit els següents criteris específics:

- Només s'anotaven aquelles persones les quals el seu nom apareixia transcrit correctament. En el cas de les persones en que el text apareixia quan no s'havia detectat cap track o que simplement no s'hagués detectat el text s'identificaven com a desconeguts. En algun cas on s'havia realitzat la trasncripicó correctament però el nom era lleugerament diferent del que apareixia en pantalla, s'identificava amb el nom de la transcripció errònia.
- Fotografies, quadres o altres imatges dins el vídeo que s'havien detectat com a falsos tracks s'anotaven amb la id de la persona real com si fossin la persona en qüestió.

Per obtenir les corbes ROC en els sistemes de verificació, donat un mètode concret s'ha realitzat un escombrat del llindar de decisió i s'ha calculat els TPR i FPR a partir de els TP, FP, TN, FN totals calculats com la suma dels resultats individuals de cada vídeo de la seqüencia d'entrenament.

4.1. Comparativa d'algoritmes de verificació de tracks

En aquesta primera prova, com s'ha comentat anteriorment, es buscava quina era la millor forma de utilitzar els vectors de característiques a l'hora de identificar tracks.

En primera instancia s'ha pogut observar que en el mètode de tots contra tots, utilitzar el sistema de classificació soft o classificació hard es irrellevant, ja que amb tots dos mètodes s'obtenen uns resultats pràcticament idèntics, la única petita diferencia entre els dos recau en el temps d'execució.

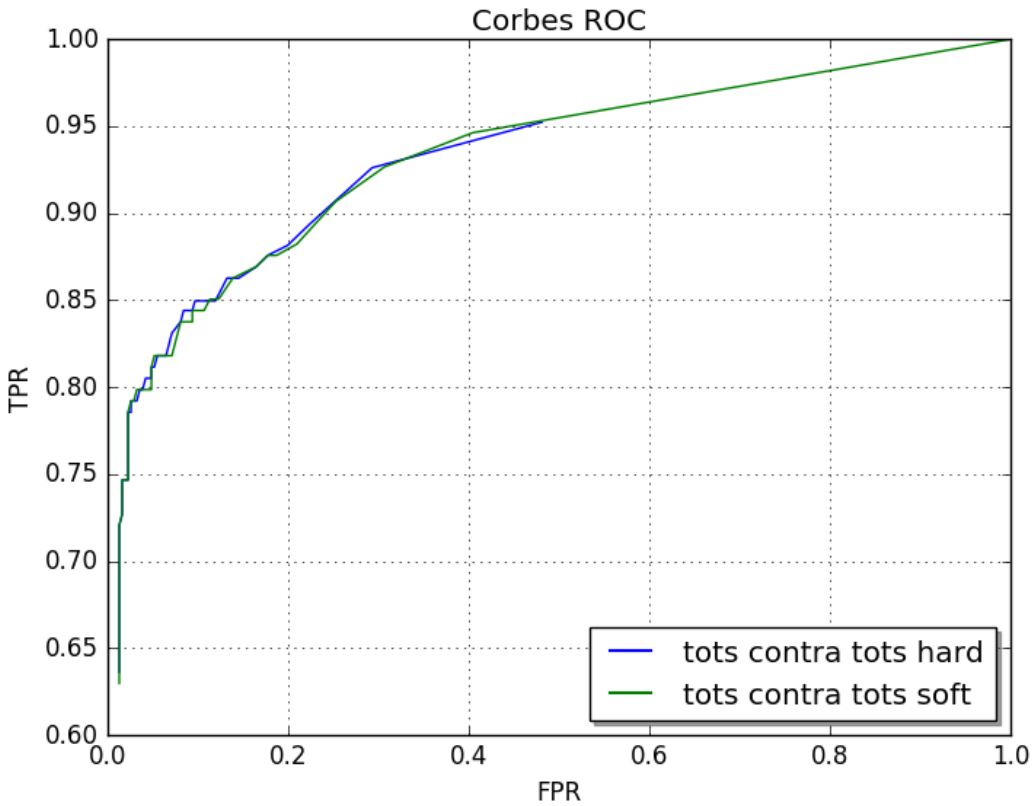


Figura 18: Comparativa de les corbes ROC obtingudes per el sistema de tots contra tots soft I tots contra tots hard. En els 2 casos, el classificador utilitzat ha estat el Naive Bayes

Si ens fixem en la resta de corbes, de seguida podem observar que el mètode del vector representatiu és el que obté pitjors resultats. El sistema de tots contra tots, tot i que lleugerament millor que l'anterior, presenta una cost computacional d'un ordre de magnitud superior als altres dos, així que el mètode de la mitja vectorial es el clar guanyador en aquest apartat.

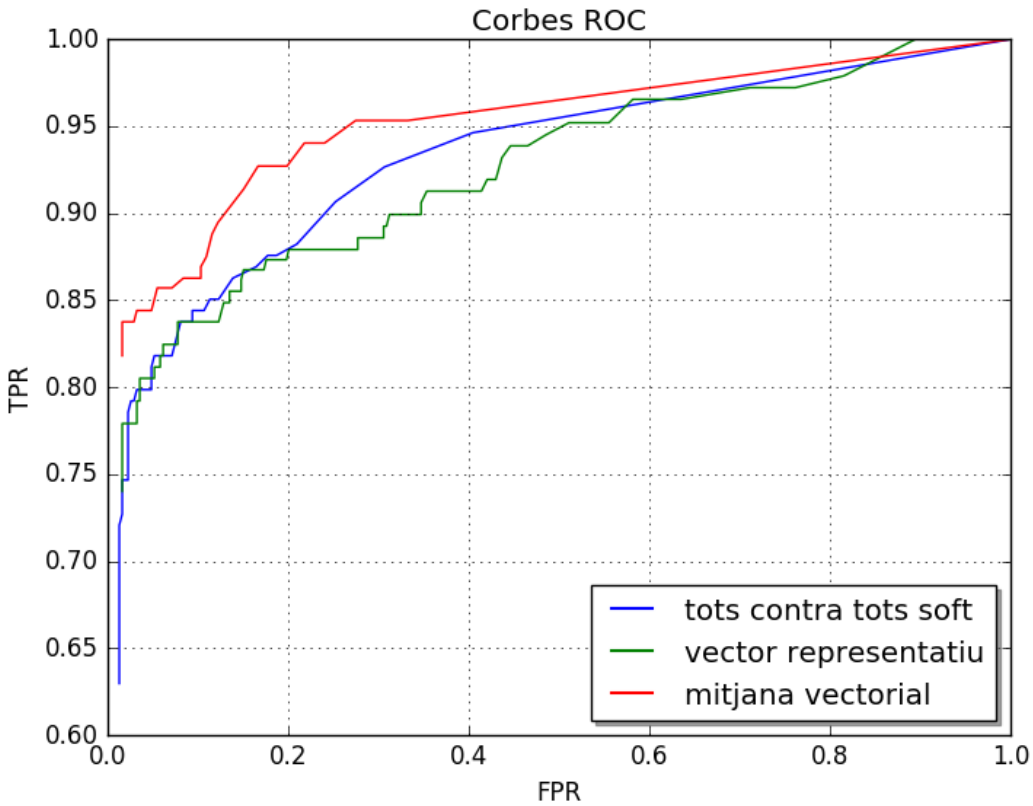


Figura 19 Comparativa de les corbes ROC dels mètode tots contra tots, vector representatiu i mitja vectorial. En tots els casos el classificador utilitzat ha estat el Naive Bayes.

4.2. Comparativa Naive Bayes amb Joint Bayes

Una vegada seleccionat el mètode de comparativa, quedava per decidir quin classificador obtindria el millors resultats. El Joint Bayes o el Naive Bayes.

Tot i que en una primera instancia podria semblar que el Joint Bayes hauria de ser millor ja que el seu rendiment teòric es superior al Naive, al final no ha resultat així. Els motius d'aquests resultats possiblement estan causats per la fase d'entrenament. El Joint bayesian requereix d'una base d'entrenament molt més gran i profunda que en el cas del Naive. Tot i que s'ha utilitzat una base de dades externa per entrenar-lo mitjançant *Transfer Learning* amb més de 500 identitats i més de 20 fotos per identitat, es possible que no haguí estat suficient ja que en article [11] utilitzaven una base de dades 3 vegades més gran i amb menys característiques per vector. A més a més, les identitats amb que s'han entrenat el joint bayesian estan totalment desvinculades de les identitats que apareixen durant els vídeos, fet que pot haver influït en el rendiment final.

També s'ha provat d'utilitzar la tècnica de PCA per tal de reduir la dimensionalitat dels vectors de característiques i així aconseguir un millor entrenament. Tot i així els resultats han resultat ser molt similars al sistema joint complert i no s'ha obtingut una millora substancial en la zona de treball

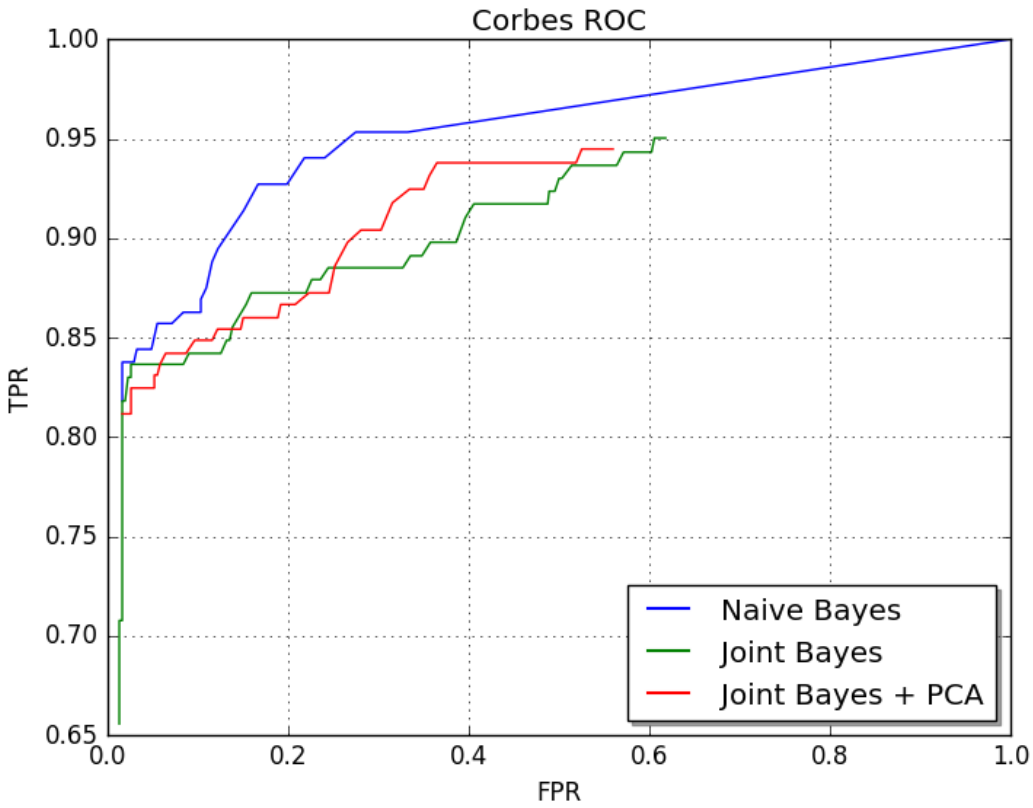


Figura 20 Comparativa de les corbes ROC entre el classificador Bayesià, el classificador Joint bayesian i el classificador joint bayesian utilitzant PCA. En tots 3 classificadors s'ha utilitzat el mètode de la mitja vectorial.

4.3. Frontaltització

Un altre dels temes pendents a resoldre era valorar si valia la pena frontalitzar les imatges o no abans de introduir-les a la xarxa convolucional. Doncs que esta demostrat que mitjançant la frontalització es pot aconseguir una lleugera reducció de de l'error en CNN [3], s'ha provat d'aplicar aquestes mateixes tècniques per intentar millorar-ne els resultats. Tot i així, en aquest projecte no s'ha disposat del mateix software de frontalizacio de manera que després de provar amb frontalització i sense, s'ha pogut observar que en el nostre cas la frontalitzacio les imatges no implica una millora de rendiment sinó tot el contrari.

Aquest resultats tenen certa lògica si tenim en compte que el frontalitzador no esta optimitzat per el conjunt de imatges que s'utilitzen i en alguns casos, el resultats de la frontalització són molt pobres. Segurament amb tècniques de frontalització més avançades com [] es podrien aconseguir resultats molt més satisfactoris.

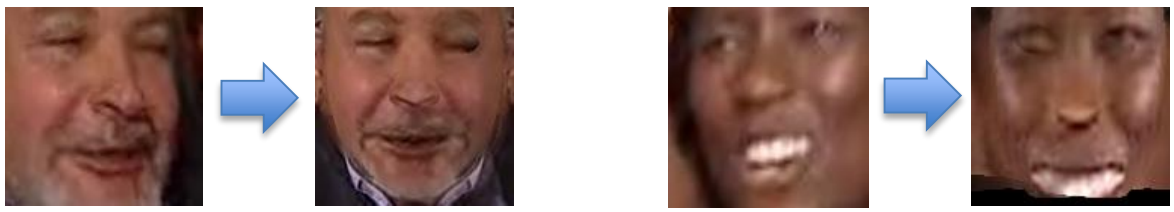


Figura 21: En aquesta figura és mostren 2 exemples del procés de frontalització mitjançant el software utilitzat en el projecte. Mentre que per algunes imatges com les de l'esquerra s'aconsegueixen molt bons resultats, en algunes altres imatges com les de la dreta els resultats són nefastos.

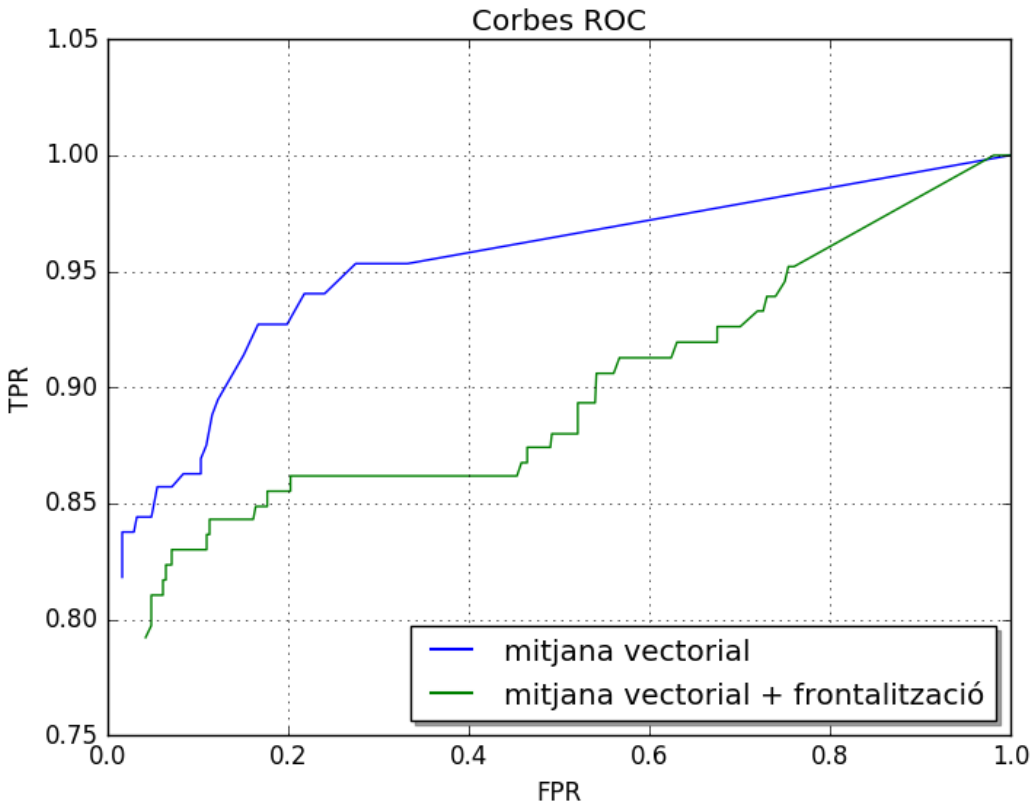


Figura 22: Comparació de les corbes ROC obtingudes mitjançant forntalització facial i sense.

4.4. Clustering

En aquest projecte s'ha decidit desenvolupar un sistema de clustering utilitzant també els vectors de característiques extrets a partir de CNN.

Per tal d'avaluar-ne el seu rendiment s'ha decidit comparar-lo tant amb el sistema de la submissió del 2015 com amb el millor mètode de verificació, la mitjana vectorial.

Mitjançant un escombrat del nombre de clústers per a un conjunt de 3 vídeos d'entrenament, podem observar que qualitativament el sistema basat en deep learning aconsegueix obtenir un rendiment lleugerament superior.

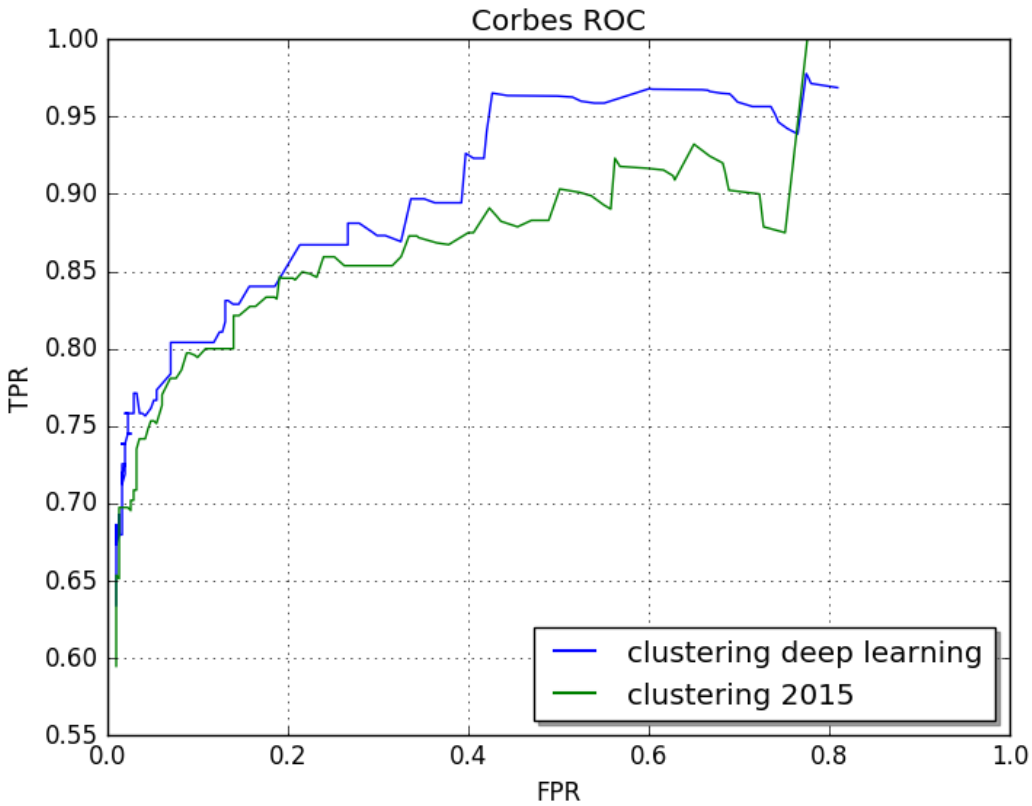
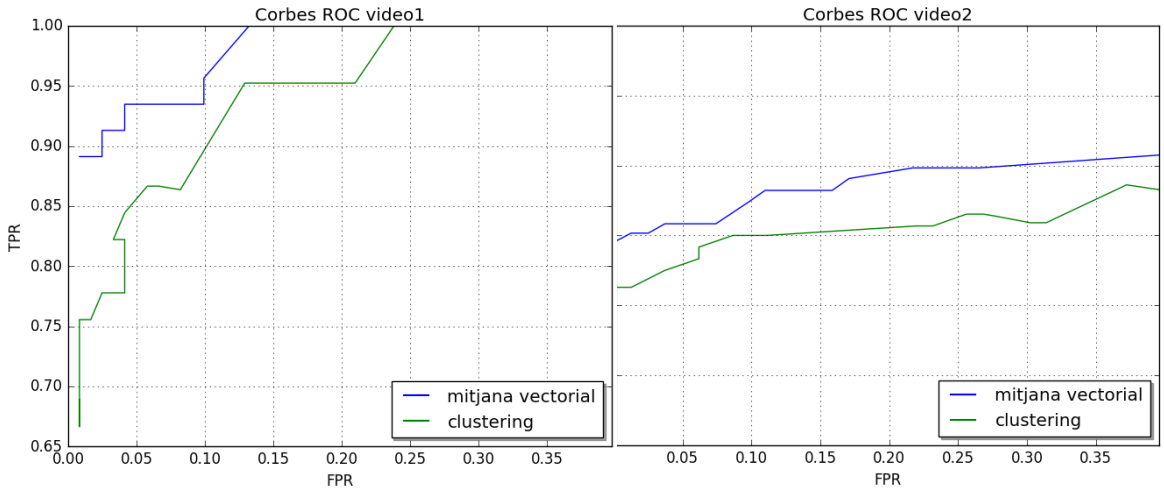


Figura 23: Comparativa entre les corbes ROC del Sistema de clustering utilitzant descriptors extrets mitjançant CNN I els descriptors del projecte Passat.

Per tal de poder comparar el sistema de clustering *deep learning* amb el millor mètode de verificació, s'han realitzat les comparacions per a cada vídeo individualment, ja que en el sistema de clustering cada vídeo utilitza el seu propi nombre de clústers, per tant, seria injust utilitzar les metodologies esmentades anteriorment.

És pot observar, que en tots 3 vídeos, en el sistema de verificació mitjançant la mitjana vectorial s'obtenen millors resultats.



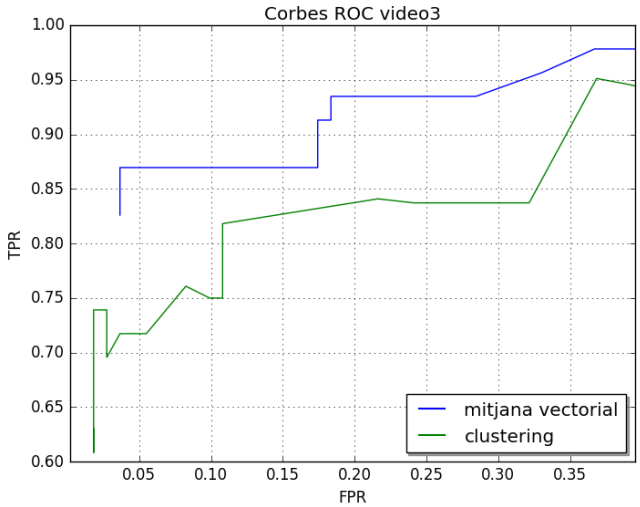


Figura 24: Comparativa de les corbes roc individuals entre el Sistema de clustering I el Sistema de verificació utilitzant promig vectorial.

4.5. Resultats d'Avaluació

Finalment, seleccionat el mètode de la mitja vectorial amb classificador Bayesia sense forntalització i fixant el llindar d'identificació òptim per la seqüencia d'entrenament (0.999998) s'han obtingut els següents resultats per els 3 vídeos d'avaluació:

A la dreta de cada cas és mostra els resultats màxims que s'haurien obtingut amb un sistema com la submissió de l'any passat.

Nom del vídeo:	FPVDB07010304	FPVDB07010705	FPVDB07010806
Nombre total Tracks	151	185	144
Nombre de TOK	32	41	33
True Positives	44/40	54/48	42/42
True Negatives	97/89	111/113	97/93
False Positives	5/14	15/13	1/5
False Negatives	5/7	5/11	4/3
Precission	0.90/0.74	0.78/0.79	0.98/0.89
Recall	0.90/0.85	0.92/0.814	0.91/0.93
Accuracy	0.93/0.86	0.89/0.87	0.97/0.94
FMeasure	0.90/0.79	0.84/0.8	0.94/0.91
MAP score	91.37/84.48	92.45/84.13	94.99/92.83

Taula 2: Comparativa dels resultats d'avaluació entre el sistema desenvolupat i la proposta de l'any passat.

La puntuació MAP del sistema anterior és en realitat és una cota superior, ja que s'ha seleccionat manualment el numero de clúster òptim en cada cas, en un sistema real, el numero de clústers es calcularia mitjançant algoritmes automàtics que empitjorarien notablement els resultats.

També és interesant mencionar que el sistema d'assignació de confiances esta més ben adaptat pel sistema de clustering que no pas pel sistema de verificació amb mitjana vectorial.

A partir de la taula 2 podem observar que mitjançant verificació deep learning s'obté una millora de resultats per sobre el 5,8%

5. Pressupost

El pressupost d'aquest projecte recau únicament en el desenvolupament ja que no s'ha realitzat cap implementació física.

La carrega de treball ha estat de 12 crèdits ETC equivalents a 300h de treball. Per aquest pressupost s'ha considerat un salari d'Enginyer junior equivalent a 15 eur/l'hora. També s'ha considerat el cost associat a la supervisor del projecte per part d'un enginyer sènior amb una dedicació d'una hora a la setmana amb un cost de 36 eur/h.

Durant el projecte s'han utilitzat diverses eines de software i serveis computacionals. De manera que s'ha decidit incorporar el cost del lloguer d'un servidor d'equivalent potencia durant 5 mesos i la compra de les diferents llicencies de software.

Amb tots els criteris anteriors, el pressupost total aproximat d'aquest treball ascendeix a 7631€ euros.

Llicencia Matlab	2000€
Lloguer de servidors	375€[1]
300h de treball a 15eur/hora	4500€
Supervisió	756€
Pressupost Total	7631€

Taula 3: Desglossament dels pressupostos

[1] Cost calculat utilitzant les tarifes de l'empresa Baltic Servers per al lloguer de 5 mesos de servidor amb GPU dedicada.

6. Conclusions:

Durant el desenvolupament d'aquest projecte s'ha pogut observar que les tècniques d'extracció de característiques basades en xarxes neuronals convolucionals proporcionen resultats superiors als descriptors facials convencionals. També s'ha comprovat que un sistema de verificació facial és molt més eficient que sistemes de clustering per a tasques d'anotació no supervisada.

Després de provar amb tots el mètodes desenvolupats s'ha arribat a la conclusió de que la combinació de verificació mitjançant mitjana vectorial, classificador gaussià, i sense etapa de frontalització obté els millors resultats. Tot i així, encara és podria aconseguir molt més rendiment si la resta del programa s'optimitzés per treballar amb aquest mètode. Ja que per exemple, l'entrenament del classificador bayesià, s'entrena mitjançant distancies de vectors individuals mentre que les distancies obtingudes en la verificació corresponen a vectors promitjats que tenen una estructura diferent. A causa d'això les distancies generades durant la verificació no es corresponen amb les distancies utilitzades amb l'entrenament i les probabilitats que genera el classificador esdevenen completament abaixades, afectant alhora el sistema de confiança.

També seria interesant continuar la recerca amb tècniques de frontalització més avançades, i provar de treballar amb un join Bayesian entrenat correctament.

De cara a un futur també es podria mirar de millorar els sistemes de pre-processat, ja que alguns dels tracks detectats erròniament (detecció facial errònia en objectes inanimats) han aparegut com a falsos positius en el sistema de verificació facial.

Bibliografia:

[1] Varas González, David, et al. "UPC system for the 2015 MediaEval multimodal person discovery in broadcast TV task." *MediaEval 2015 Multimedia Benchmark Workshop*. 2015.

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[3] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." *British Machine Vision Conference*. Vol. 1. No. 3. 2015.

[4] Houghton, Ricky. "Named faces: Putting names to faces." *IEEE Intelligent Systems and their Applications* 14.5 (1999): 45-50.

[5] Satoh, Shin'ichi, Yuichi Nakamura, and Takeo Kanade. "Name-it: Naming and detecting faces in news videos." *Ieee Multimedia* 6.1 (1999): 22-35.

[6] Everingham, Mark, Josef Sivic, and Andrew Zisserman. "Hello! My name is... Buffy"--Automatic Naming of Characters in TV Video." *BMVC*. Vol. 2. No. 4. 2006.

[7] Canseco-Rodriguez, Leonardo, Lori Lamel, and Jean-Luc Gauvain. "Speaker diarization from speech transcripts." ICSLP, 2004.

[8] Canseco, Leonardo, Lori Lamel, and J-L. Gauvain. "A comparative study using manual and automatic transcriptions for diarization." *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*. IEEE, 2005.

[9] Bredin, Hervé, et al. "Person Instance Graphs for Named Speaker Identification in TV Broadcast." *Proceedings of Odyssey*. 2014.

[10] scikit-learn developers, «http://scikit-learn.org/stable/modules/naive_bayes.html

[11] Chen, Dong, et al. "Bayesian face revisited: A joint formulation." *European Conference on Computer Vision*. Springer Berlin Heidelberg, 2012.

[12] Wikipedia contributors. "Hierarchical clustering." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 2 Jun. 2016. Web. 26 Jun. 2016.

[13] Rouseeuw, P. J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*. Vol. 20, No. 1, 1987, pp. 53–65.

[14] Calinski, T., and J. Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics*. Vol. 3, No. 1, 1974, pp. 1–27.

[15] Hassner, Tal, et al. "Effective face frontalization in unconstrained images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[16] Farag, Amal, et al. "A Bottom-up Approach for Pancreas Segmentation using Cascaded Superpixels and (Deep) Image Patch Labeling." *arXiv preprint arXiv:1505.06236* (2015).

[17] Wikipedia contributors. "Information retrieval." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 16 Jun. 2016. Web. 27 Jun. 2016.

Glossari

CNN: *Convolutional Neural Network*, terminologia anglesa per referir-se a les capes neuronals convolucionals.

MAP: *Mean Average Precision*, mètrica utilitzada per mesurar la qualitat d'anotació.

MAP: *Maximum a posteriori estimation*, criteri utilitzat per realitzar estimacions probabilístiques.

PCA: Analisis de components principals, tècnica utilitzada per reduir la dimensionalitat de un conjunt de dades.

ROC: *Receiver operating characteristic*, representació gràfica de la sensibilitat enfront la especificat.

TP: Positiu verdader

TN: Negatiu verdader

FP: Fals positiu

FN: Fals negatiu

TPR: *False positive rate*

FPR: *False negative rate*

TOK: Conjunt de tracks que han estat identificats mitjançant el solapament temporal de noms escrits.