UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

IDENTIFICACIÓ NO-SUPERVISADA DE PERSONES EN PROGRAMES DE TELEVISIÓ

Treball de Fi de Grau

Presentat a la

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

per

Anna Martí Aguilera

En compliment parcial

dels requisits per al grau en

Enginyeria de Sistemes Audiovisuals

Supervisor: Josep Ramon Morros Rubió

Barcelona, Juny 2016

Abstract

The enormous amount of visual data generated nowadays creates a strong need for annotation tools to enable search and retrieval of information present in the videos. One of the most relevant information is the identity of people.

The aim of this project is to implement non-supervised algorithms of text and face recognition, to identify relevant people appearing in Broadcast TV. This project achieves avoiding manual annotations with an automatic annotation system.

Resum

La enorme quantitat de dades visuals que es genera avui en dia crea una forta necessitat de obtenir tècniques d'anotació per a poder realitzar cerques d'informació en els vídeos. Una de la informació més rellevant és la identitat de les persones.

L'objectiu d'aquest projecte és proposar uns algorismes no supervisats de reconeixement facial i de text per a la identificació de les persones en les transmissions de TV, per obtenir un sistema d'anotació de vídeo automàtic, i evitar així les anotacions manuals.

Resumen

La enorme cantidad de datos visuales generados hoy en día crea una fuerte necesidad de obtener técnicas de anotación para poder realizar búsquedas de información en los vídeos. Una de la información más relevante es la identidad de las personas.

El objetivo de este proyecto es proponer unos algoritmos no supervisados de reconocimiento facial y de texto para la identificación de las personas en transmisiones de TV, para obtener un sistema de anotación de vídeo automático y así evitar las anotaciones manuales.

Agraïments

Primer de tot m'agradaria agrair al tutor d'aquest projecte, en Josep Ramon Morros Rubió per fer-me de guia i per tots els consells donats, m'han estat de gran ajuda per poder desenvolupar dia rere dia aquest projecte i fer-lo créixer, sense ell no hagués sigut possible.

També volia agrair a l'Albert Gil per donar suport en l'ús del servidor i per ajudar a solucionar els problemes que anaven sorgint.

Voldria agrair també als companys per el suport que ens hem donat durant el transcurs de la carrera i sobretot a la Belen Luque, per totes les hores compartides fent cadascuna el seu projecte donant-nos suport i ànims quan el necessitàvem.

Finalment, agrair a la meva família per la confiança que han tingut en mi i per l'esforç que han fet perquè pugui arribar fins a on he arribat.

Historial de revisions registre d'aprovacions

Revisió	Data	Objectiu
0	31/05/2016	Creació del document
1	08-10/06/2016	Revisió del document
2	12-16/06/2016	Revisió del document
3	18-19/06/2016	Revisió del document
4	21-26/06/2016	Revisió del document

LLISTA DE DISTRIBUCIÓ DEL DOCUMENT

Nom	Correu electrònic
Anna Martí Aguilera	annamartiaguilera@gmail.com
Josep Ramon Morros Rubió	ramon.morros@upc.edu

Escrit per:		Revisat i aprovat per:	
Data	31/05/2016	Data	27/06/2016
Nom	Anna Martí Aguilera	Nom	Josep Ramon Morros Rubió
Rol	Autor del projecte	Rol	Supervisor del projecte

Taula de Continguts

Llistat de Il·lustracions

Llistat de Taules:

1. Introducció

1.1. Contextualització del projecte

Les bases de dades d'arxius de televisió estan creixent ràpidament en grandària. La necessitat d'aplicacions que facilitin la cerca d'aquests arxius ha portat als investigadors a dedicar esforços per al desenvolupament de tecnologies que creen índexs.

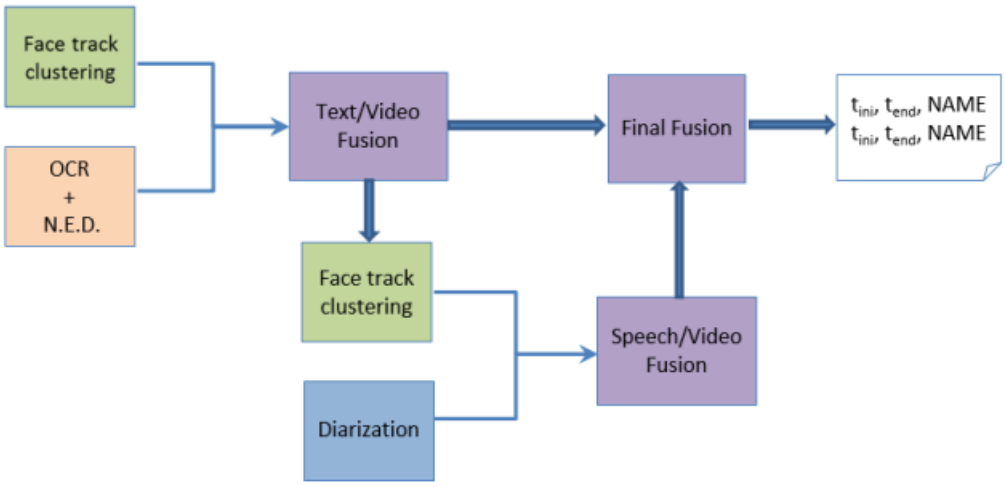
Des de la UPC es participa en un projecte europeu anomenat Camomile[1] que té com a finalitat millorar els sistemes d'anotació. Els membres del projecte organitzen i participen en les avaluacions MediaEval[2], en la tasca Multimodal Person Discovery in Broadcast TV Task.

MediaEval és una iniciativa de referència dedicada a l'avaluació de nous algoritmes per l'accés i la recuperació multimèdia. Es centra en els aspectes socials i humans de les tasques multimèdia.

Aquesta tasca proposada a MediaEval representa una extensió del desafiament francès REPÈRE[3] ja finalitzat, que es centrava en el reconeixement multimodal de persones en emissions de TV, la diferència és que, en la tasca de MediaEval, només s'admet l'ús d'algoritmes no-supervisats, és a dir, només s'admeten algoritmes que no es basin en etiquetes o models preexistents.

Aquesta tasca va ser proposada per primera vegada en el 2015 i la UPC va participar-hi. Aquest any s'hi presenta de nou amb l'objectiu d'assolir millors resultats respecte l'any anterior.

Aquest treball es basa en el desenvolupament de dos blocs dins del diagrama de blocs del projecte: Face track clustering i OCR + N.E.D.



Il·lustració 1 Diagrama de blocs del projecte desenvolupat a la UPC.

[1] http://www.chistera.eu/projects/camomile
[2] http://multimediaeval.org/mediaeval2016/
[3] http://www.defi-repere.fr/index.php?id=27&L=1

1.2. Declaració d'objectius

L'objectiu d'aquest projecte és bàsicament col·laborar en el desenvolupament d'un sistema d'anotació automàtica de vídeos de programes de televisió mitjançant tècniques de reconeixement tant de cares com de text i implementant un sistema de extracció de noms de persones.

1.3. Requeriments i Especificacions

Requeriments del projecte:

- Dissenyar i implementar algoritmes per a l'agrupació facial en vídeos de televisió
- Dissenyar i implementar algoritmes per a l'agrupació de text i extracció de noms de persones en vídeos de televisió

Especificacions del projecte:

- Avaluar els algoritmes implementats en el corpus de MediaEval 2015
- Utilitzar la mètrica d'avaluació *Mean Average Precision* (MAP) per avaluar els resultats
- Pel desenvolupament dels algoritmes, s'utilitzaran els llenguatges de programació següents: Python, C++ i Matlab.

1.4. Pla de treball

1.4.1. Work Packages

WP1. Proposta de projecte i redacció del pla de treball

WP2. Lectura de l'estat de l'art

WP3. Desenvolupament d'un detector facial

WP4. Desenvolupament d'un detector de text

WP5. Desenvolupament d'un Named Entity Detection

WP6. Revisió Crítica del Treball

WP7. Avaluació dels resultats i proposta de millores

WP8. Redacció de la memòria del Treball

1.4.2. Justificació

En la proposta del pla de treball, l'organització del desenvolupament de cada part està basada en el resultats obtinguts l'any anterior i per tant, prioritza aquells blocs que tenien resultats més fluixos, en aquest cas: detecció facial, detecció de text, i extracció de noms de persones. Es desestima el desenvolupament de tècniques de seguiment facial fins que tots els blocs no estiguin acabats, per tant, en tot cas es desenvoluparia en el WP7- Avaluació de resultats i proposta de millores.

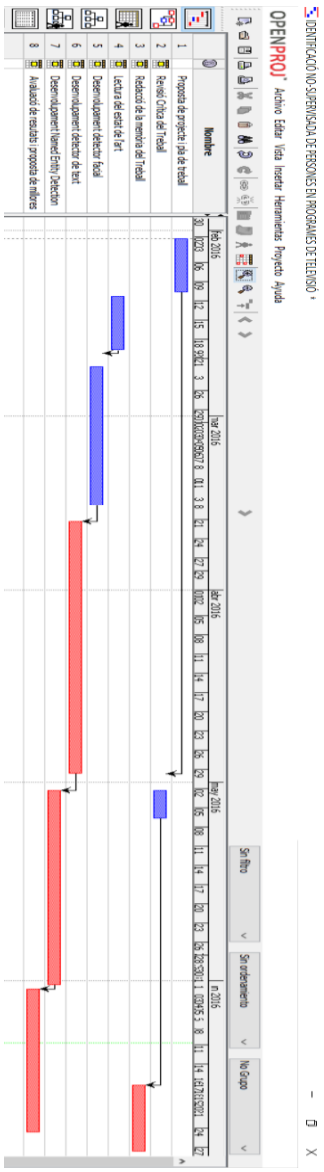
1.4.3. Incidències i desviacions respecte al pla de treball original

El pla de treball original repartia equitativament el temps de dedicació en ambdós blocs principals. Un cop començat el treball, ràpidament es va trobar una solució per implementar el sistema que apliqués un detector facial en vídeo, i vam decidir tirar endavant i començar a desenvolupar el sistema de detecció de text.

El temps dedicat a la implementació del sistema de detecció de text ha sigut molt més del planificat originalment, degut a que les tecnologies existents s'apliquen i donen molt bons resultats en imatges quan es pot determinar en quins llocs de la imatge apareixen els noms de persones. Quan els noms poden aparèixer en qualsevol lloc de la imatge, es fa complicat diferenciar els noms de persona del text que apareix en la imatge.

1.4.4. Diagrama de Gantt

A continuació es presenta el diagrama del Gantt del projecte:



Il·lustració 2: Diagrama de Gantt del Projecte

2. Estat de l'art de la tecnologia aplicada en aquest treball

Des que es va proposar el desafiament francès REPERE l'any 2011, s'ha avançat molt en la investigació de la tecnologia d'anotació automàtica per vídeo.

Per abordar aquest problema, els projectes d'investigació s'han centrat en resoldre dos preguntes: "Qui apareix en els vídeos?" i "Qui està parlant?"

En aquest treball només ens centrarem en respondre la pregunta "Qui apareix en el vídeo?".

Aquest treball es basa en tècniques d'aprenentatge no supervisat per al desenvolupament dels algoritmes.

L'aprenentatge no supervisat és un tipus d'algoritme d'aprenentatge automàtic que s'utilitza per treure conclusions a partir de conjunts de dades sense etiquetar.

El mètode més comú d'aprenentatge no supervisat és l'anàlisi clúster, que s'utilitza per a l'anàlisi de dades exploratòries per trobar patrons ocults o agrupació en les dades. Els clústers es modelen mitjançant una mesura de similitud que es defineix en mètriques com ara la distancia euclidiana o la probabilística.

2.1. Detecció i Seguiment de Cares

2.1.1. Detecció facial

La detecció facial pot considerar-se un cas específic de la detecció d'objectes. En la detecció d'objectes, la tasca consisteix en trobar en una imatge la localització i la mida dels objectes que pertanyin a una classe en concret.

Els algoritmes de detecció facial identifiquen característiques facials extraient *landmaks* – que són uns punts concrets que reconeixen les cares de la imatge – de la cara d'un subjecte.

Per la detecció facial hem estudiat dues implementacions:

- Viola-Jones de OpenCV [1]
- Detector *Histogram of Oriented Gradients* [2] de la llibreria Dlib[3].

2.1.1.1. Viola-Jones

Aquest algoritme consta de 3 etapes:

- Selecció de característiques *Haar*
- Entrenament *Adaboost*
- Classificadors en cascada

Selecció de característiques *Haar*:

Les característiques sol·licitades pel marc de detecció universal involucren les sumes de píxels de la imatge dins de les àrees rectangulars. Com a tals, tenen certa semblança amb funcions de base Haar, que s'han utilitzat prèviament en el camp de detecció

d'objectes basat en imatges. No obstant això, ja que les característiques utilitzades per Viola i Jones es basen en més d'una àrea rectangular, són generalment més complexes.

- Característiques *Haar.*
  Tots els rostres humans comparteixen algunes propietats similars. Aquestes regularitats poden ser identificades mitjançant característiques Haar. En les dues imatges següents es mostren dos exemples:

Il·lustració 3: Característica Haar que s'assembla al pont del nas aplicada sobre la cara. Font: Wikipedia Viola–Jones object detection framework

Il·lustració 4: Característica Haar que s'assembla a la zona dels ulls i les galtes aplicada sobre la cara. Font: Wikipedia Viola–Jones object detection framework

Entrenament *AdaBoost*:

La velocitat amb la qual es poden avaluar les característiques no compensa adequadament pel seu nombre, però, per exemple, en una sub-finestra estàndard de 24x24 píxels, hi ha un total de M=162,336 possibles característiques, i seria prohibitivament car avaluar-les totes quan es prova una imatge.

Per tant, el marc de detecció d'objectes empra una variant de l'algoritme d'aprenentatge AdaBoost tant per seleccionar les millors característiques com per capacitar els classificadors que les utilitzen. Aquest algoritme construeix un classificador "fort" com a combinació lineal de classificadors ponderats "febles".

Classificació en cascada:

- Un simple classificador de dues característiques pot arribar a la taxa de detecció de gairebé el 100% amb una taxa de Falsos Positius d'un 50%. Aquest classificador pot actuar com una primera etapa d'una sèrie de classificadors per eliminar les finestres més negatives.
- Una segona etapa amb 10 característiques pot fer front a les finestres negatives supervivents en la primera etapa.
- Una cascada de classificadors gradualment més complexes aconsegueix resultats encara millors. Tot i que la avaluació dels classificadors més forts generats pel procés d'aprenentatge es pot fer ràpid, és insuficient per executar-ho en temps real. Per aquesta raó, els classificadors forts estan disposats en una cascada en ordre de complexitat, on cada classificador successiu està entrenat només amb aquelles mostres seleccionades que passen a través dels classificadors anteriors. El primer classificador a la cascada fa servir només dos característiques per aconseguir reduir a gairebé a la meitat el nombre de vegades que s'avalua tota la cascada.

Els avantatges d'aquesta tècnica són:

- Còmput de característiques extremadament ràpid
- En lloc d'escalar les imatges, escala les característiques
- Fa una selecció de característiques eficient.

Els desavantatges són:

- El detector és més eficaç en les imatges de cares frontals
- Difícilment pot fer front als 45 ° de rotació cara tant al voltant de l'eix vertical com de l'horitzontal
- Sensible a les condicions d'il·luminació
- Podríem obtenir múltiples deteccions de la mateixa cara a causa de sub-finestres superposades.

2.1.1.2. Detector Histogram of Oriented Gradients

El detector Histogram of Oriented gradients, detector HOG, es basa en avaluar histogrames locals normalitzats de les orientacions del gradient de la imatge. La idea bàsica és que la aparença i la forma d'un objecte local pot estar ben caracteritzada per la distribució de gradients locals d'intensitat, fins i tot sense el coneixement precís de les posicions del gradient. A la pràctica, s'implementa dividint la imatge en petites regions anomenades cel·les, i per cada cel·la, s'acumula un histograma local de gradients orientats unidimensional sobre els píxels de la cel·la, cada bin del histograma conté un rang de orientacions en valor absolut del gradient – si tenim 2 bins, un bin contindrà els gradients orientats entre 0-90º, i l'altre, els gradients orientats entre 90-180º.

Per tal de normalitzar les respostes locals, s'acumula la mesura d'un histograma local d'energia sobre unes regions formades per varies cel·les anomenades blocs, cada cel·la es normalitza amb el valor del bloc que la conté. Aquests blocs normalitzats s'anomenen descriptors HOG.

El càlcul del descriptor HOG es duu a terme sobre tota la finestra de detecció. Un cop tenim tots els vectors de característiques HOG, es passen com a entrada a un classificador lineal SVM. La finestra de detecció s'escaneja a través de la imatge en totes les posicions i escales.

Aquest sistema s'implementa amb els següents paràmetres:

- Espai de color RGB sense correcció gamma
- Filtre de gradient sense suavizat [-1 0 1]
- Histograma de 9 bins en 0-180º
- Mida de les cel·les: 8x8 píxels
- Mida del bloc: 16x16 píxels (4x4 cel·les)
- Finestra de detecció 64x128

2.1.2 Seguiment de cares

El seguiment d'objectes (*object tracking*) s'ha estudiat extensivament en el context de Visió per Computadors a causa de les moltes aplicacions visuals – com per exemple els robots autònoms, la videovigilància o el seguiment de cares humanes – que utilitzen els algoritmes de seguiment d'objectes. El seguiment d'objectes en situacions complexes ha

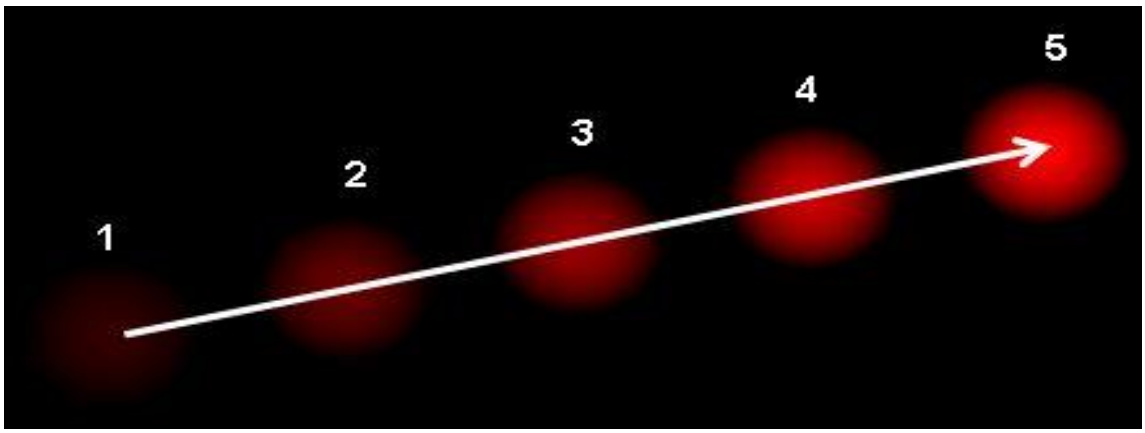
de fer front a la incertesa i al error. Hi ha varies tècniques que intenten posar solució al problema, que es podrien classificar en 2 categories:

- **Tracking basat en flux òptic.** En aquesta categoria estudiarem el mètode de tracking basat en la imatge amb l'*optical flow* de Lucas-Kanade [4].

- **Tracking basat en filtratge**. En aquesta categoria s'estudien els filtres de partícules. La idea bàsica del filtre de partícules és l'aproximació de la densitat posterior mitjançant un filtre bayesià recursiu usant un conjunt de partícules amb pesos assignats.

Optical Flow de Lucas-Kanade

El flux òptic és el patró de moviment aparent dels objectes d'imatge entre dos frames consecutius causats pel moviment de l'objecte o la càmera. És un camp vectorial 2D on cada vector és un vector de desplaçament que mostra el moviment dels punts del primer frame al segon.



Il·lustració 5: Vector de flux òptic d'un objecte en moviment en una seqüència de vídeo. Font: Wikipedia Optical Flow

La imatge anterior mostra una pilota en moviment en 5 frames consecutius. La fletxa mostra el vector de desplaçament.

El mètode de Lucas-Kanade treballa sobre diverses suposicions:

- Tots els canvis en els valors dels píxels son deguts al moviment, no a canvis d'il·luminació
- Tots els píxels al frame t es poden relacionar amb algun píxel del frame $t + \Delta t$ (i viceversa)
- El moviment és petit
- Hi ha coherència en el moviment, els píxels tenen un moviment similar al dels seus veïns.

Amb aquestes suposicions, es pot estimar el moviment amb una aproximació de Taylor de primer ordre, però només és vàlid en aquestes condicions. A la pràctica, el moviment no és petit, per tant, la solució és disminuir la resolució de la imatge per reduir la

magnitud del moviment. Això ens porta a un enfocament multi-resolució de la tècnica Lucas-Kanade.

En aquest enfocament, la imatge es va delmant fins a reduir la magnitud del moviment, i es calcula LK en el nivell més alt. Els vectors resultant s'interpolen i es compensa la referencia en el següent nivell. Es torna a calcular LK en el següent nivell i així successivament.

El seguiment d'objectes es pot fer de dues formes:

a) Mitjançant el seguiment dels píxels que conformen l'objecte

b) Fent una detecció de les cares en tots els frames de la seqüència i relacionant les deteccions mitjançant els vectors del flux òptic (tracking by detection)

Filtres de Partícules

La idea bàsica del filtre de partícules és l'aproximació de la densitat posterior usant un filtre bayesià recursiu basat en un conjunt de partícules amb pesos assignats. Per a cada frame d'una seqüència d'imatges en el marc de seguiment visual, un filtre de partícules en general consta de tres passos:

- Mostreig
- Ponderació
- Selecció

S'extreu un conjunt de partícules de la distribució proposades en el pas de mostreig. En l'etapa de ponderació, cada partícula es pondera basant-se en la relació de la seva veritable probabilitat amb la seva probabilitat aproximada utilitzant la distribució proposada. A continuació i finalment es seleccionen les partícules (re-mostrejades) d'acord amb la densitat posterior estimada per obtenir una distribució de pes uniforme en l'etapa de selecció.

2.2. Reconeixement de text i identificació de noms de persones

Una de les tècniques utilitzades per identificar els noms de les persones que apareixen en el vídeo és la transcripció de la parla i la extracció dels noms de persones pronunciats pels parlants. Aquesta tècnica d'extracció de noms no l'hem contemplat en aquest projecte perquè suposa molts falsos positius -no sempre qui és anomenat en un programa hi apareix- i els resultats són molt sorollosos.

En aquest projecte, per identificar els noms de les persones que apareixen en el vídeo, ens basarem en els noms escrits en la pantalla. D'aquesta manera ens assegurem que les persones que finalment etiquetem en el vídeo són rellevants i hi apareixen.

2.2.1 Reconeixement del text

Existeixen varies eines de reconeixement de text, i totes les més importants utilitzen la tècnica *Optical Character Recognition* (OCR) per al reconeixement de caràcters en imatges.

Les tècniques de detecció de text contemplades en aquest treball són les següents:

- *LOOV* per Johann Poignant, Laurent Besacier, Georges Quenot, Franck Thollard [5]
- *Reading Text in the Wild* per Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman [6]
- Class-specific Extremal Regions for Scene Text Detection [7]
- Detecció i segmentació de text a través de l'estimació de l'ample de traç i BPT [8]
- Region-Based Caption Text Extraction [9]

2.2.1.1. LOOV

Presenten un sistema OCR de vídeo que detecta i reconeix textos superposats en el vídeo. Realitza una adaptació de les imatges per a un sistema OCR estàndard, i fa un post-processament final per combinar múltiples transcripcions de la mateixa caixa de text. Aquesta tècnica es basa purament en la informació que dóna la imatge.

S'utilitza la següent estratègia per a la detecció del text:

En primer lloc, troba un màxim de candidats, i en segon lloc, descarta els textos no pertinents.

- Aplica un filtre de Sobel vertical i horitzontal amb la finalitat de detectar el contorn dels caràcters. Una dilatació amb poques iteracions permet connectar caràcters consecutius.
- A continuació s'aplica una operació d'obertura per aïllar els components connexos; i es detecten les línies de text amb un mètode basat en les projeccions horitzontals i verticals. Això genera una gran quantitat de falsos positius.
- S'utilitza un segon dispositiu de detecció local amb el refinament de l'aprenentatge automàtic per filtrar aquestes falses alarmes.
- Un filtrat temporal s'aplica pel vídeo per poder fer un seguiment temporal del text. Aquest pas aprofita el fet que un text donat apareix de forma idèntica en molts quadres successius. Aquesta informació temporal s'utilitza per a filtrar els falsos positius i a més, per recuperar les caixes en que la detecció falla a nivell local.

Un cop detectada la localització del text en la imatge, i comprovat que té consistència temporal, es realitza el pas de la extracció del text. Per poder-ho dur a terme, s'ha d'adaptar el quadre de text al software OCR. S'augmenta artificialment la resolució de les imatges mitjançant una interpolació bi-cúbica. A continuació s'aplica una binarització de la imatge utilitzant un llindar calculat amb l'algoritme Sauvola [10]. Per millorar la qualitat de la transcripció del text, s'aplica OCR en varies imatges per una mateixa caixa.

2.2.1.2. Reading Text in the Wild:

La implementació d'aquest Sistema segueix els següents passos:

- Generació de propostes de *Bounding boxes* de paraules + Filtratge de Propostes
- Reconeixement del text
- Combinació

Aquest procés segueix vagament la separació detecció/reconeixement – una etapa de detecció de paraules seguit d'una de reconeixement de paraules. No obstant això, aquestes dues etapes no estan molt diferenciades, ja que s'utilitza la informació proporcionada pel reconeixement de paraules per a combinar i puntuar els resultats de la detecció al final, donant lloc a un sistema integrat de divisió de text més fort.

La etapa de detecció està basada en mètodes de detecció dèbils però molt ràpids per generar les propostes de Bounding boxes de paraules. Utilitzar propostes de regions evita la complexitat computacional de la recerca exhaustiva amb finestra lliscant. Degut a la gran quantitat de propostes Falsos Positius, s'utilitza un classificador *random forest* per a filtrar el nombre de propostes fins a una mida manejable.

La segona etapa d'aquest framework proporciona un resultat de reconeixement de text per a cada proposta generada en la etapa de detecció. Amb un enfocament per al reconeixement de paraules completes, es proporciona com a entrada de una xarxa neuronal convolucional tota la regió de la paraula detectada. Presenten un model de diccionari el qual planteja la tasca de reconeixement com a una tasca de classificació multi-via a través d'un diccionari de 90.000 possibles paraules.

Finalment, s'utilitza la informació obtinguda del reconeixement de text per a actualitzar els resultats de les deteccions.

2.2.1.3. Class-specific Extremal Regions for Scene Text Detection

Aquest algoritme es basa en seleccionar Extremal Regions (ERs) adequades en tot l'arbre de components de la imatge. La selecció de les ER adequades es fa mitjançant un classificador seqüencial entrenat per a la detecció de caràcters.

L'arbre de components de la imatge es construeix posant un llindar per un valor creixent a cada pas des de 0 a 255, i unint els components connectats de successius nivells en la jerarquia per la seva relació d'inclusió

L'arbre de components pot contenir un gran nombre de regions. Per tal de seleccionar de manera eficient regions adequades entre tots els ERs, l'algoritme fa ús d'un classificador seqüencial amb dues etapes diferenciades.

En la primera etapa es calculen uns descriptors computacionalment simples per a cada regió r i s'utilitzen com a característiques d'un classificador que estima la probabilitat condicional de la classe p(r|caràcter). Només es seleccionen els ERs que corresponguin a un màxim local de la probabilitat condicionada p(r|caràcter).

En la segona etapa, els ERs que han passat la primera etapa es classifiquen en classes caràcter/no-caràcter utilitzant característiques que donen més informació però alhora són computacionalment més complexes.

Després del filtratge de ERs, els candidats de caràcters s'agrupen en blocs de text d'alt nivell (és a dir, paraules, línies de text, paràgrafs, ...).

2.2.1.4. Detecció i segmentació de text a través de l'estimació de l'ample de traç i BPT

Aquesta tècnica es basa en l'estimació d'amplada de traç com a característica distintiva dels caràcters per a la detecció de text i en el creixement de regions per a la segmentació.

L'algoritme consta de 6 etapes principals:

- Detecció de contorns mitjançant el detector de contorns Canny.
- Transformada *Stroke Width Transform*, SWT, per a obtenir l'amplada de traç.
- Segmentació *Binary Partition Tree*, BPT, per a obtenir els components candidats a caràcters.
- Filtratge de components basada en les característiques d'amplada de traç i geometria.
- Creació de cadenes amb el propòsit d'unir els candidats a caràcter per a obtenir una línia de text candidata.
- Binarització de la imatge per a obtenir el text, mitjançant la informació de la localització del text candidat donada al pas anterior.

2.2.1.5. Region-Based Caption Text Extraction

Aquesta tècnica es concentra en la detecció de text de subtítols per a imatges fixes. El text de subtítols es refereix a text afegit a l'interior d'un quadre rectangular, d'alt contrast en relació amb el fons i amb aspecte de textura. Tots els algoritmes utilitzen un mateix model de imatge basat en regions jeràrquic.

Aquesta tècnica té 3 etapes:

- Text candidate spotting
- Text characteristic verification
- Tree analysis

En la primera etapa s'estimen les característiques de textura usant un anàlisis wavelet. La segona etapa, la verificació de característiques, es basa en característiques geomètriques les quals s'estimen explotant el model de la imatge basat en regions. L'anàlisi de la jerarquia de regió proporciona els objectes de text finals. El pas final d'anàlisi de consistència per a la sortida es realitza mitjançant un algoritme de binarització que estima robustament els llindars en l'àrea de suport del text.

2.2.2 Extracció de noms

Per la extracció de noms, s'utilitza el que es diu Named Entity Recognition (NER), es tracta d'un anàlisis lingüístic i classificació del text automàtic.

Les implementacions que hem contemplat per al desenvolupament d'aquest projecte son dues:

- Stanford NER [11]
- Freeling [12]

2.2.2.1 Stanford NER

El Stanford NER es una implementació del *Named Entity Recognizer*. Etiqueta seqüències de paraules en un text. Genera etiquetes per a 3 classes: *PERSON, ORGANIZATION* i *LOCATION*. Està preparat per a ser utilitzar en textos en anglès, però també suporta altres idiomes.

La majoria dels models estadístics utilitzats actualment en el processament del llenguatge natural, representen només l'estructura local, una limitació clau en moltes tasques, ja que el llenguatge natural conté una gran quantitat d'estructura no local. Aquesta implementació de Stanford es diferència per utilitzar un mètode general per a la solució d'aquest problema, la substitució d'algoritmes d'inferència aproximats permetent d'aquesta manera tractar amb models amb estructura no local. L'algoritme que utilitzen s'anomena mostreig de Gibbs, un simple algoritme de Montecarlo que és apropiat per a la inferència en qualsevol model probabilístic factoritzat. Tot i que el mostreig de Gibbs és àmpliament utilitzat en altres llocs, hi ha hagut molt poc ús d'ell en processament del llenguatge natural. Aquí, la fan servir per afegir dependències que no són locals per seqüenciar models per a l'extracció d'informació.

2.2.2.2 Freeling

Presenten un conjunt d'analitzadors lingüístics bàsics. Permet les següents funcions:

- Divisió de frases i *tokenization*
- Anàlisis morfològic
- Reconeixement de múltiples paraules
- Detecció de noms propis
- Reconeixement d'expressions numèriques
- Etiquetatge *Part-of-Speech* (PoS) seguint un model trigrama HMM

L'arquitectura del sistema es basa en dos tipus d'objectes: objectes de dades lingüístiques i objectes de processament.

Característiques:

Aquesta eina inclou les funcionalitats descrites anteriorment per a Espanyol, Català i Anglès. Els diccionaris morfològics consisteixen en:

- Anglès: Més de 160.000 formes d'uns 78.000 lemes, obtinguts a partir de WSJ. Les 200 formes més freqüents han estat revisades a mà. El diccionari pot contenir una mica de soroll.
- Espanyol: Prop de 71.000 formes, conté tots els lemes de la classe tancat més els 5.000 lemes més freqüents de classe oberta, codificats a mà.

- Català: Prop de 46.000 formes, conté tots els lemes de la classe tancat més els 5.000 lemes de classe oberta més freqüents, codificats a mà.

S'espera que l'analitzador morfològic cobreixi tots els *tokens* de categoria tancada i més del 80% de *tokens* de text sense restriccions de la categoria oberta. No obstant això, les paraules desconegudes es gestionen a través de probabilitats condicionals d'etiquetes PoS, es proposen les etiquetes PoS més probables per a cada paraula no inclosa en el diccionari morfològic.

L'etiquetador PoS ofereix una precisió superior al 95% per tots els idiomes. El sistema és capaç d'analitzar morfològicament un text a una velocitat propera a les 6000 paraules/segon en un processador P4 2.8GHz.

Aquestes dades les proporcionen els desenvolupadors de la eina Freeling en l'article "FreeLing: An Open-Source Suite of Language Analyzers" [8] realitzat pels autors.

3. Metodologia / Desenvolupament del projecte:

3.1. Base de dades:

La base de dades d'entrenament consta del corpus "6 Months of Broadcast News" de INA – el que va servir de conjunt de test en l'edició del 2015. Consta de 110 hores de vídeo.

La base de dades de test d'aquest any consta de 3 corpus diferents:

- Un subconjunt de vídeos de INA fet a partir d'una setmana sencera de 3 canals de televisió Francesos. Consta de 370 hores de vídeo.
- Un subconjunt del corpus DW/EUMSSI. Consta de 60 hores de vídeo.
- Un subconjunt de vídeos de l'emissora catalana 3-24. Consta de 90 hores de vídeo.

Il·lustració 6: Imatge vídeo Corpus 6 Months of Broadcast News

Il·lustració 7: Imatge d'un dels vídeos del corpus DW/EUMSSI

Il·lustració 8: Imatge d'un vídeo del corpus INA

Il·lustració 9: Imatge d'un dels vídeos de 3-24

3.2. Detecció facial:

El projecte presentat l'any passat per la UPC, en la detecció facial s'utilitzava el mètode Viola-Jones de OpenCV. Donava bons resultats però també tenia molts falsos positius. Per solucionar-ho es pretén utilitzar la tècnica de detecció facial amb HOG de la llibreria dlib, descoberta durant la lectura de l'estat de l'art.

Aquesta llibreria anomenada dlib, té un detector facial que presumeix de tenir les mateixes deteccions positives i moltes menys falses deteccions que la llibreria OpenCV.

Combinant el detector facial de la llibreria dlib i la implementació del seguiment de cares de l'any anterior – basat en el mètode de Lucas-Kanade mencionat en l'estat de l'art – podem comparar els dos detectors avaluant els resultats a nivell de track.

Fent l'avaluació d'ambdues tècniques sobre 5 vídeos etiquetats manualment, i fent la mitjana de tots els vídeos s'obtenen els següents resultats:

	OpenCV	**Dlib**
Tacks no detectats	10	6
Tracks detectats correctament	76	80
Falses deteccions	7	1

Taula 1: Comparativa dels detectors de OpenCV i de Dlib sobre 5 videos

Amb la llibreria OpenCV obtenim un valor de precisió del 0,91, i amb la llibreria Dlib obtindríem una precisió de 0,98. El valor de la precisió indica el rati entre el nombre de deteccions correctes respecte al nombre de deteccions (tant correctes com incorrectes).

Amb la llibreria OpenCV obtenim un valor de Recall de 0,88 i amb la llibreria Dlib obtenim un valor de 0,93. El valor de Recall indica el percentatge de deteccions correctes respecte al total de elements que haurien de estar etiquetats.

Amb aquestes dades, decidim seguir endavant amb el nou detector facial.

Com s'ha explicat en el pla de treball, la millora del seguiment facial no es va considerar prioritària ja que d'entrada dóna bons resultats, s'utilitza el tracking basat en Lucas Kanade, tracking by detection mencionat a l'estat de l'art. Per tant, es tanca el bloc de detecció facial i es deixa pendent a possibles millores la tasca de desenvolupament d'una nova tècnica de seguiment temporal de cares.

3.3. Detecció de Text:

Per la detecció de text, inicialment es prova el Tesseract, que utilitza un detector molt simple basat en binarització de la imatge. Com que aquesta eina només binaritza la imatge d'entrada, es preprocessa la imatge amb un filtre de Sobel seguit d'un *adaptive thresholding* abans de passar el Tesseract. Com que les imatges utilitzades només tenen text superposat, aquesta tècnica dona uns resultats bastant dolents.

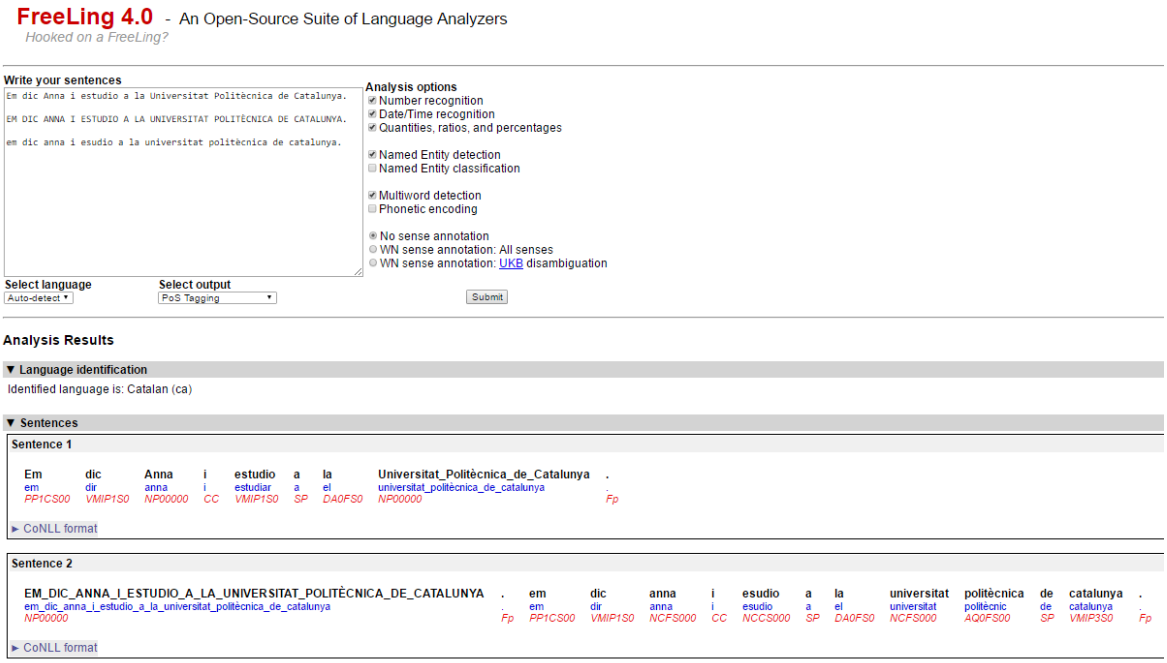
Amb la lectura del estat del art es descobreixen varies eines de detecció de text, però es va decidir utilitzar la eina LOOV per raons de velocitat de procés. Aquesta utilitza el OCR després d'aplicar un preprocessament de la imatge. A més, està orientada a vídeo perquè per cada frame localitza la posició del text i fa un seguiment temporal d'aquest.

En vistes dels resultats obtinguts pot observar que s'adapta a les necessitats del projecte: tot i que en el text extret per la eina LOOV hi ha molts falsos positius - interpreta formes en la imatge que no són text i les escriu com si ho fos-, no afecta als resultats del projecte ja que d'aquest text, es filtrarà tot el que no son noms propis.

3.4. Extracció de noms:

Un cop el text s'ha detectat, s'ha de extreure el que son noms de persones, i quedar-se amb la informació temporal que dóna el LOOV.

Per fer-ho es prova la eina Freeling, que té la mateixa funcionalitat que el NER de Stanford però amb l'avantatge que també permet analitzar text en català. Quan es comencen a fer proves ens n'adonem que tots els resultats els classifica com a noms propis, i pensem que el fet que el text estigui tot en majúscules pot ser el motiu per a que això passi. Per assegurar-nos preguntem als desenvolupadors del codi, expliquem el nostre cas on el text que tenim està escrit tot en majúscules i que si això podria afectar en la classificació de les paraules. Ens confirmen que aquesta eina en part es recolza en la informació que donen les majúscules i les minúscules per saber si una paraula és un nom propi. Podem veure un exemple on provem la mateixa frase en 3 casos diferents en la Il·lustració 10.

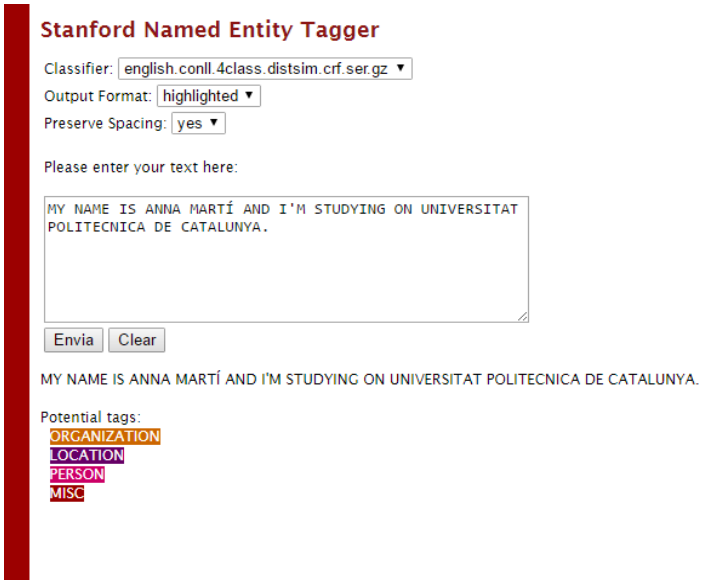


Il·lustració 10: Captura de Pantalla de la demo online de la eina Freeling (http://nlp.lsi.upc.edu/freeling/demo/demo.php)

En el primer cas, el text està escrit com ho estaria en un document, amb els signes de puntuació adequats i amb les majúscules i minúscules utilitzades correctament.

En el segon cas, veiem dos exemples diferents: La primera frase que està tota escrita en majúscules la detecta sencera com si fos un sol nom, i la segona frase que està escrita en minúscules la detecta correctament, excepte els noms.

Provem el classificador Stanford NER i ens trobem amb el mateix problema, com es pot veure en la Il·lustració 11.



Il·lustració 11: Captura de Pantalla de la demo online del Stanford NER. http://nlp.stanford.edu:8080/ner/process

Al estar el text en majúscules es perd la informació que dóna una majúscula al mig d'una frase, per tant, el sistema és incapaç de detectar el que és un nom en aquestes condicions.

Veient que les tècniques d'anàlisis lingüístic no es poden utilitzar en el context del projecte, s'intenta buscar una solució alternativa. Com que l'anàlisi lingüístic era necessari únicament per filtrar allò que etiquetava com a Persona, es prova de generar un diccionari de noms i cognoms, creant un llistat amb els noms i un altre amb els cognoms existents a cada país dels vídeos en la base de dades: Alemanya, França i Catalunya. Per fer-ho, es busca un llistat amb els noms dels residents al país per exemple extraient la informació del cens.

Per dur-ho a terme es crea un sistema que fa que cada línia de sortida del extractor de text passi per un bloc de normalització del text i per un altre que prengui la decisió a partir d'uns requisits en concret.

3.4.1. Normalització del Text:

Es desenvolupa un bloc anomenat *text_normalization* on s'eliminen o es substitueixen tots els caràcters que poden induir a l'error, per exemple: si la eina d'extracció de text extreu el text però no detecta un accent, al comparar el nom extret amb el nom amb accent del llistat, no el detectaria; en el cas contrari, si la imatge tingués una línia a prop del text i amb la extracció de text ho detectés com un accent, podríem perdre aquesta detecció. La solució d'aquest problema implica que tant el text de les llistes com el de la sortida del extractor de text estiguin normalitzats de la mateixa manera. Per normalitzar fem dues coses:

- S'eliminen tots els accents.
- Es converteixen els caràcters no-ASCII en el seu equivalent en ASCII, per exemple: es substitueix el caràcter 'ç' pel caràcter 'c'.

3.4.2. Bloc de Decisió:

Per acceptar com a nom una línia de text s'han de complir uns requisits:

- La línia de text ha de tenir com a mínim 2 paraules i com a màxim 5
- Com a mínim ha d'aparèixer una paraula de la línia de text al llistat de noms i una de diferent al llistat de cognoms
- Com a mínim la línia de text ha de tenir 5 caràcters, sense contemplar l'espai.
- Posem un llindar de paraules, que implica que el número de paraules dins les llistes respecte al número de paraules totals ha de ser superior o igual a 0.75:
    - Si la línia té 5 paraules, 4 han d'estar dins les llistes. (donem un marge d'error d'una paraula per si el detector de text ha fallat)
    - Si la línia té 4 paraules, 3 han d'estar dins les llistes. (donem un marge d'error d'una paraula per si el detector de text ha fallat)
    - Si la línia té 2 o 3 paraules, totes han d'estar dins les llistes. Si donéssim marge al detector per fallar, donaríem peu a que resultés en molts falsos positius.

Un cop normalitzat el text, s'aconsegueix pràcticament una detecció del 100% dels noms, però també s'obtenen molts falsos positius, per exemple, hi ha noms a les llistes que son monosíl·labs molt fàcils de detectar per error amb la detecció de text -per exemple 'il' es pot detectar per error si en la imatge apareixen dues barres petites- si aquesta línia de text compleix els requisits exposats anteriorment, aquesta paraula que no és un nom es detecta com a tal. Per solucionar-ho es proposen dues restriccions addicionals:

Restricció 1. Es desenvolupa un *Heat-Map* de totes les deteccions que passen el filtre de nom i s'eliminen aquells que apareguin menys de 100 vegades (menys de 4 segons de vídeo):

Un *Heat-Map* és un mapa de calor, partint d'una imatge amb tots els píxels a 0 (de color negre) es construeix sumant 1 als píxels en les posicions on es detecten noms. Com que els vídeos que s'utilitzen són vídeos de programes de televisió, la localització dels noms de persones es concreta en un espai determinat. Això fa que sigui fàcil eliminar tots els noms detectats en zones poc habituals que el més probable és que siguin Falsos Positius. Un cop fet el *Heat-Map*, es binaritza el resultat per obtenir una imatge representativa de la localització del text durant tot el transcurs del vídeo: una imatge de fons negre i quadres blancs on apareix text amb freqüència. En les il·lustracions 13, 14 i 15 podem veure un exemple de les imatges resultants en dos programes diferents.

Il·lustració 13: Heat-Map sense binaritzar

Il·lustració 12: Heat-map per un programa de Institut national de l'audiovisuel (INA)

Il·lustració 14: Heat-map per un vídeo del programa 3-24 de Televisio de Catalunya (TVC)

Això serveix per eliminar tots els noms espuris que podrien resultar d'una mala detecció, considerant noms espuris tots aquells que estiguin en una localització la qual aparegui menys de 100 vegades en el transcurs del vídeo. Per eliminar-los, es mira el valor dels píxels de la localització del text en el Heat-Map, si aquest valor es inferior a 100, aquest text es descarta del procés de extracció de noms.

Restricció 2. Filtrar els noms a la llista de noms que siguin preposicions mono-sil·làbiques del tipus: 'Carlos de la Fuente' eliminant les paraules 'de' i 'la' de la llista de noms i eliminant-les de la línia de text a la hora de filtrar -perquè no sigui aquest el problema de no superar el llindar del 0.75 de paraules en la llista:

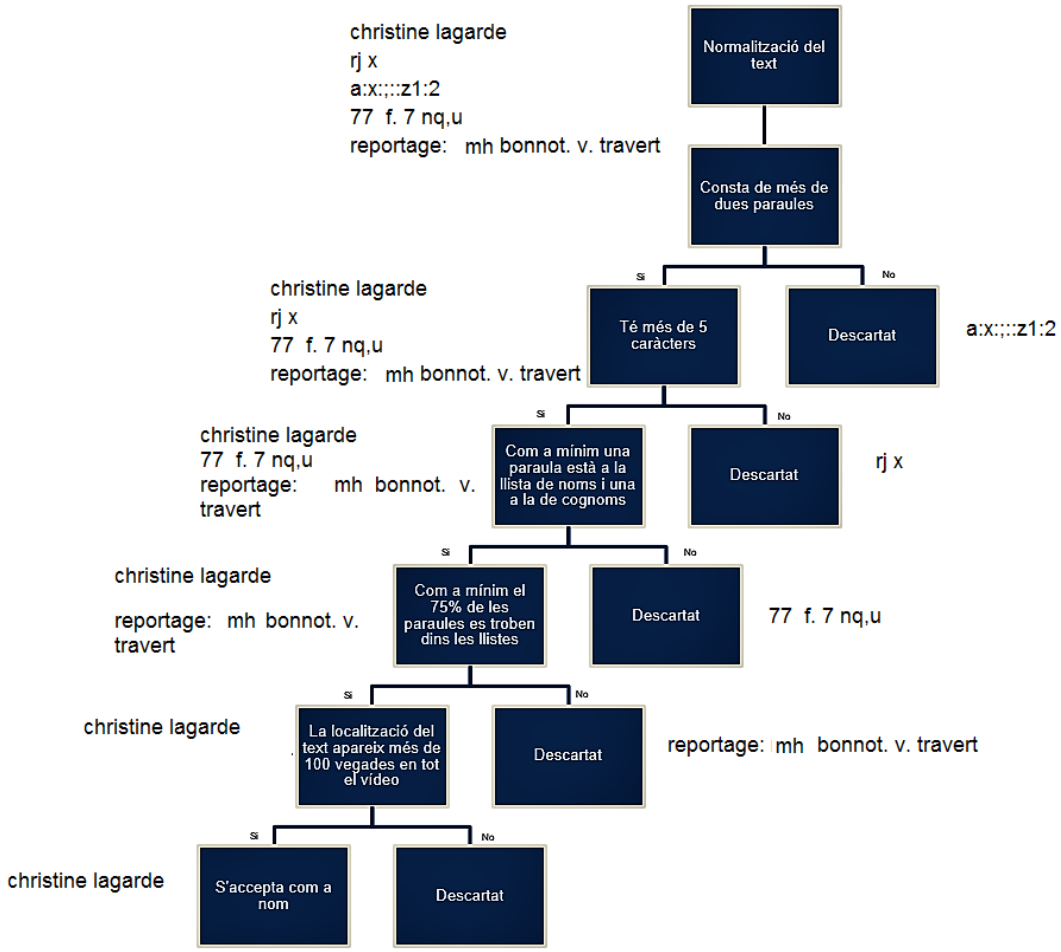
Per a les preposicions que precedeixen els noms, s'ha creat una llista de noms 'neutrals' que no ponderen ni positivament ni negativament per a la detecció, simplement es descarten per evitar errors. Per fer-ho, s'han eliminat els noms de dues paraules de les llistes de noms i cognoms i s'han apuntat a la llista de noms 'neutrals' tots aquells que precedeixin a noms. Som conscients que això pot eliminar alguna bona detecció, però analitzant-ho creiem que és molt millor deixar d'etiquetar molts falsos positius amb el cost de perdre un nom real monosil·làbic, que en percentatge respecte a tots els noms existents és molt petit.

Aquesta restricció s'aplica al inici del bloc d'adaptació del text explicat anteriorment.

En la Il·lustració 16 es mostra un exemple de com funciona el extractor de noms per unes línies de la sortida del detector de text mostrades en la Taula 2.

Frames	Bounding boxes $[y_i, x_i, y_f, x_f]$	Text
7051-7154	[219, 153, 233, 304]	CHRISTINE LAGARDE
7058-7067	[140, 81, 165, 193]	RJ x
7060-7081	[82, 279, 89, 347]	a;x:;::x1:2
7070-7108	[106, 283, 114, 304]	77 ' f .7 'nq-,Ü
54007-54073	[220, 117, 232, 304]	REPORTAGE : MH BONNOT. V. TRAVERT

Taula 2: Línies de sortida del detector de text



Il·lustració 15: Blocs de restriccions i exemple de funcionament del bloc

3. Resultats

La mètrica d'avaluació utilitzada és el *Mean Average Precision,* MAP. La mètrica MAP genera un valor únic que resumeix el rendiment d'un sistema.

El Precision and Recall és una mètrica de valor únic en base a tota la llista de documents retornats pel sistema. Per als sistemes que retornen una llista ordenada de documents, com és el nostre cas, és convenient tenir en compte l'ordre en que es presenten els documents retornats. Average Precision (AP) calcula el valor mitjà de Precision en l'interval des de Recall=0 fins a Recall=1. El MAP, per tant, és la mitjana dels valors AP en totes les consultes.

Hi ha dues tècniques de detecció facial per crear Tracks i dues tècniques de detecció de text:

A. Bloc de detecció facial i seguiment temporal del baseline, que utilitza el mètode de Viola-Jones per al detector facial i el Tracking by Detection de Lucas-Kanade per al seguiment facial.
B. Bloc de detecció facial i seguiment temporal desenvolupat en aquest projecte, que utilitza el detector HOG-SVM per la detecció facial i el Tracking by Detection de Lucas-Kanade per al seguiment facial.
C. Bloc de detecció de text i extracció de noms del baseline, que utilitza el mètode LOOV per a la detecció de text en una finestra molt reduïda, sense extracció de noms.
D. Bloc de detecció de text i extracció de noms desenvolupat en aquest projecte, que utilitza la eina LOOV per a la detecció de text i que es basa en diccionaris de noms propis per a la extracció de noms.

S'estudien les diferents combinacions dels sistemes d'extracció de tracks (A,B) amb els de detecció de text (C,D).

Calculant el MAP(A, C) obtenim els resultats del baseline, els quals podrem utilitzar per a comparar amb la resta. Si el comparem amb el resultat de MAP(A,D) podrem obtenir informació de com funciona el detector de text desenvolupat en el projecte. De la mateixa manera, comparant MAP(A,C) amb el resultat de MAP(B,C) podrem obtenir informació de com funciona el detector facial. Finalment, si comparem MAP(A,C) amb el resultat de calcular MAP(B,D), podrem comparar el funcionament d'ambdós sistemes.

El resultat mostrat en la Taula 3 és el resultat global de 3 vídeos de la base de dades INA de 2015.

MAP(A,C)	49.04 %
MAP(A,D)	32.97 %
MAP(B,C)	72.15 %
MAP(B,D)	66.98 %

Taula 3: Comparativa de resultats amb el base-line

Com es podia esperar, el resultat del detector de text no es millor que el del baseline, això és degut a que en el baseline els noms sempre es cercaven en la mateixa localització, una barra estreta situada en la part inferior de la imatge, on tot el text que apareixia eren noms. En el sistema actual, el text pot aparèixer en qualsevol part de la imatge, on també pot aparèixer text que no correspon a noms de persona.

L'extracció de noms s'ha fet a partir d'uns diccionaris de noms creats per nosaltres mateixos amb informació extreta d'internet, aquests diccionaris han estat revisats ja que inicialment hi havia noms que hem descartat – per exemple, noms de dues consonants, però segurament queda molta per revisar ja que aquesta feina s'ha fet manualment.

Veiem que la eina utilitzada per fer la detecció i el seguiment facial funciona millor que el del baseline, i que el detector de text funcionarà amb qualsevol format de programa de televisió, no només en aquells on els noms de persona estiguin escrits en la part inferior de la pantalla. En la base de dades de vídeos d'aquest any, el text apareix en localitzacions diferents i per tant, el sistema del baseline no ens serveix.

4. Pressupost

Hores dedicades al desenvolupament del projecte : 580

Work Package	Hores dedicades
Proposta de projecte i redacció del pla de treball	25
Lectura de l'estat de l'art	40
Desenvolupament d'un detector facial	60
Desenvolupament d'un detector de text	120
Desenvolupament d'un Named Entity Detection	140
Revisió Crítica del Treball	15
Avaluació dels resultats i proposta de millores	110
Redacció de la memòria del Treball	70

Hores dedicades a la supervisió del projecte: 1h/setmana durant 22 setmanes, 22 hores.

Tenint en compte que el projecte s'ha desenvolupat i executat en el servidor del grup de imatge, s'ha d'estimar el cost del lloguer del servidor. Considerant que el servidor llogat té les següents característiques:

	vCPU	*ECU*	*Memory (GiB)*	*Instance Storage (GB)*	*Linux/UNIX Usage*
g2.2xlarge	*8*	*26*	*15*	*60 SSD*	*0,70€ per Hour*

Fent una estimació de 400 hores de temps de procés, el cost que tindria el lloguer del servidor seria de 280 €.

Considerant que la dedicació d'un enginyer junior es paga a 10€/hora, i que les hores de supervisió del projecte es paguen a 40€/hora, preu d'enginyer sènior, i considerant les despeses en el lloguer dels servidor, aquest projecte pressupostaria amb un valor de 6960€ (5800 € + 880 € + 280 €).

5. Conclusions i futur desenvolupament:

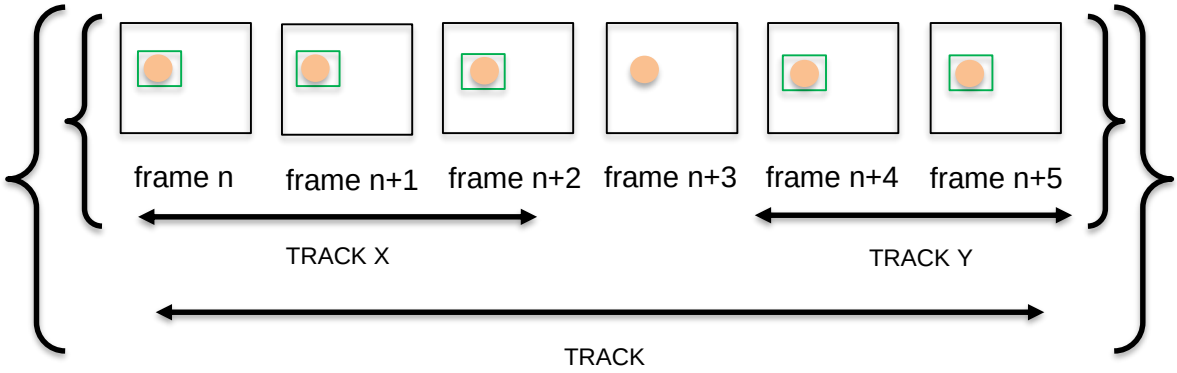
L'objectiu d'aquest projecte era desenvolupar uns algoritmes de reconeixement facial, de reconeixement de text i d'extracció de noms de persones per col·laborar en la creació d'un sistema d'anotació automàtica de vídeos de televisió, partint del projecte existent de l'any anterior.

Com es pot veure en els resultats, els blocs elaborats en aquest projecte són competents, ja que, encara que el bloc de text tingui resultats més pobres que el del baseline de 2015, és més robust respecte qualsevol canvi en la localització del text.

Aquest sistema s'ha utilitzat per genera les anotacions automàtiques que s'han distribuït com a baseline de la tasca "Multimodal Person Discovery in Broadcast TV" de MediaEval 2016.

Durant el transcurs del projecte, anaven sorgint idees de possibles millores, però no era viable desenvolupar-les en el marge de temps establert, per tant, es plantegen com a possibles millores de cara al futur desenvolupament.

- Revisar els diccionaris de noms i cognoms incloent tots aquells que no estan a les llistes.
- Provar les tècniques de detecció de text mencionades en el estat de l'art i utilitzar la més competitiva en el sentit de resultats i de temps d'execució.
- Millorar el bloc de seguiment temporal de cares fent que un cop detectats tots els *tracks*, uneixi els que, a causa d'un error en la detecció d'una cara, estiguin comptabilitzats com a dos *tracks* diferents tot i que es refereixin a la mateixa cara ja que els separa només un frame i els *Bounding boxes* estan en la mateixa localització en la imatge.



Bibliografia:

[1] Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*, *4*.

[2] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE.

[3] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*(Jul), 1755-1758.

[4] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. IJCAI, 81:674–679, 1981.

[5] Poignant, J., Besacier, L., Quénot, G., & Thollard, F. (2012, July). From text detection in videos to person identification. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on* (pp. 854-859). IEEE.

[6] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, *116*(1), 1-20.

[7] Neumann, L., & Matas, J. (2012, June). Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3538-3545). IEEE.

[8] Lladó Gargallo, J. (2014). Detecció i segmentació de text a través de l'estimació de l'ample de traç i BPT.

[9] Leon, M., Vilaplana, V., Gasull, A., & Marques, F. (2013). Region-based caption text extraction. In *Analysis, Retrieval and Delivery of Multimedia Content* (pp. 21-36). Springer New York.

[10] Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization.*Pattern recognition*, *33*(2), 225-236.

[11] Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In LREC.

[12] Poignant, J., Bredin, H., & Barras, C. (2015). Multimodal person discovery in broadcast tv at mediaeval 2015. *Proceedings of MediaEval*.

[13] Poignant, J., Bredin, H., & Barras, C. (2015). Multimodal person discovery in broadcast tv at mediaeval 2015. *Proceedings of MediaEval*.