Universitat Oberta de Catalunya (UOC)

Màster Universitari en Ciència de Dades

Treball final de màster

Àrea: Mineria de dades i machine learning

Detecció d'objectes a seqüències de vídeo

Autor: Joan Bonnín Hernández

Tutor: Gabriel Moyà Alcover

Professor: Jordi Casas Roma

Palma, 9 de juny de 2019

FITXA DEL TREBALL FINAL

Títol del treball:	Detecció d'objectes a seqüències de vídeo
Nom de l'autor:	Joan Bonnín Hernández
Nom del col·laborador/a docent:	Gabriel Moyà Alcover
Nom del PRA:	Jordi Casas Roma
Data d'entrega:	09/06/2019
Titulació o programa:	Màster universitari en Ciència de Dades
Àrea del Treball Final:	Mineria de dades i *Machine Learning*
Idioma del treball:	Català
Paraules clau:	Computer vision, Machine Learning, Object tracking

Abstract

0.1 Abstract (*English*)

The Master's Thesis project consists in the identification, selection and evaluation of different methods and systems for solving two current problems in the computer's vision field: object detection and object tracking.

To solve both tasks, we've studied classical solutions with a well-known good performance and the latest approaches based on machine learning and deep learning.

In order to make a comparison between models, a set of experiments has been done. Those experiments are built over the dataset of MOTChallenge, specifically 2017 edition. For the detection task the studied models are: DPM, SDP, Mask-RCNN and YOLOv3. For the tracking task the studied models are: CamShift, correlation filters and SORT.

The combination between different systems to solve both tasks, aims to the fact we actually have the required techniques to automatize of the tasks. Even that, the characteristics of the images to process directly affect the results' quality. To sum up, we define the best models for general scenes, but it's crystal clear that there exists the need of evaluating the context and characteristics of the scene to decide which model to use.

Keywords: Computer vision, Machine learning, Video tracking, Artificial intelligence, Object detection, Object tracking

0.2 Resum

El projecte de Treball Final consisteix en la identificació, selecció i avaluació de diferents mètodes i sistemes per a la resolució de dos problemes vigents en el camp de la visió per computador: la detecció i el seguiment d'objectes. Per resoldre ambdues tasques s'han estudiat tant solucions clàssiques amb bon rendiment, com les darreres novetats basades en aprenentatge automàtic i aprenentatge profund.

Per poder realitzar la comparativa de models, es realitzen una sèrie d'experiments. Aquests experiments es realitzen sobre el conjunt de dades del MOTChallenge, en concret a l'edició del 2017. Per la detecció s'estudien els models DPM, SDP, Mask-RCNN i YOLOv3, mentre que pel seguiment s'estudien CamShift, filtres de correlació i SORT.

La combinació de diferents sistemes per resoldre les dues tasques de forma combinada conclou que es disposen de tècniques amb bondats suficients per a l'automatització de la tasca, tot i que les característiques de les imatges a processar afecten directament a la qualitat del resultat. Tot plegat, es defineixen els millors models per escenes generals, però queda patent la necessitat d'avaluar el context i natura de les imatges a tractar per realitzar una correcta selecció i aplicació de models de detecció i seguiment.

Paraules clau: Visió per computador, Aprenentatge automàtic, Intel·ligència artificial, Detecció d'objectes, Seguiment d'objectes.

Agraïments

Vull agrair a tot el professorat que, al llarg dels anys, m'ha impulsat a ser una persona curiosa i a millorar constantment. En aquesta ocasió, és inevitable agrair-li en especial a en Biel que, d'una manera o altra, m'ha acompanyat i ajudat en tota la meva etapa com alumne de ciències de computació. Una etapa que ara es tanca, però que de ben segur es tornarà a obrir.

També agraesc la paciència i suport de na Maria, la meva companya. I les rialles que em regalen sempre "ses que ho guanyen tot", els que bufen amb vent de Llebeig i els companys de feina que s'han convertit en molt més que això.

Índex

Índex de figures

Índex de taules

Capítol 1

Proposta inicial

1.1 Descripció i justificació de la proposta

En l'actualitat ens trobem en una situació de creixement d'aplicacions directes dels camps de visió per computador i intel·ligència artificial. Dins d'aquests camps, les aplicacions amb major popularitat són els sistemes de conducció automàtica o assistida, així com la producció en cadena de productes a l'hora de separar elements de la cadena. Tot i això, aquestes tècniques es fan servir en molts altres àmbits com el mèdic, la videovigilància o fins i tot les xarxes socials.

Més concretament, aquestes dues àrees de coneixement es poden fusionar en tècniques de detecció i seguiment d'objectes. A més de l'ús per la conducció autònoma, aquestes tecnologies de seguiment es poden fer servir per identificar els fluxos de circulació (tant de vehicles com de persones) en un moment donat, per predir conglomeracions o detectar punts d'alta afluència.

L'objectiu del projecte és la comprensió i valoració de les propostes més vigents en la comunitat científica, a més de l'obtenció del coneixement necessari per desenvolupar un sistema en funció de diferents necessitats possibles. Conceptualment, la tasca d'aquests sistemes serà: una vegada detectats els elements rellevants, analitzar els moviments o interaccions entre objectes.

Aquesta aplicació, no té per què estar relacionada directament amb seqüències de circulació, sinó que pretén apropar-se a una solució més generalista. Un possible ús pràctic d'aquesta tecnologia és la generació de descripcions automàtiques de vídeos, molt útil en diferents aplicacions concretes: generació automàtica de sinopsis, descripció de les imatges per a invidents, etc.

Pel desenvolupament del projecte es fan servir diferents eines i tècniques de l'estat de l'art relatives als camps de visió per computador i aprenentatge automàtic.

1.2 Motivació personal

La motivació principal per afrontar aquest projecte té tres vessants:

La primera d'elles, treballar en un projecte més complex i complet que els vists fins al moment al llarg del màster universitari. Això em permetrà entendre millor els problemes i les casuístiques d'un problema real, que hom pot trobar en l'àmbit professional.

La segona està relacionada amb l'àmbit de la investigació. Tota la meva trajectòria professional ha estat lligada a l'empresa privada, desenvolupant productes a mesura per satisfer diferents necessitats de negoci concretes. Em sembla molt interessant deslligar-me d'aquesta forma de treballar per adquirir noves perspectives i punts de vista.

Finalment, la visió per computador és una àrea tècnica que sempre m'ha interessat i no he pogut explotar en detall durant les diferents assignatures cursades. Espero poder combinar els coneixements adquirits sobre *machine larning* i mineria de dades sobre aquesta forma de dades no estructurades que són les imatges.

1.3 Objectius del projecte

El projecte pretén assolir diferents objectius per tal de resoldre la problemàtica descrita:

- Identificar quins models permeten, en l'actualitat, identificar i segmentar diferents elements dins una imatge (*estat de l'art*).
- Definir quins són els elements rellevants a cada un dels fotogrames de la seqüència d'entrada.
- Relacionar els elements detectats a cada un dels fotogrames per tal d'identificar-ne l'evolució temporal.
- Analitzar els resultats obtinguts per diferents combinacions de models, per tal de maximitzar la bondat del sistema, ajustant bé o models o paràmetres.
- Identificar o adaptar el millor sistema, segons els resultats de l'estudi, per a la resolució del problema descrit.

Tot plegat, existeixen altres objectius transversals com ara l'obtenció de coneixement sobre les diferents àrees o la comprensió del procés de recerca en un projecte de mineria de dades i *machine learning*.

1.4 Metodologia

Per al desenvolupament del projecte es proposa una metodologia iterativa. En lloc d'usar un model en cascada, es treballarà en diferents cicles basats en prototipus que dependran dels resultats anteriors, per tal de garantir una evolució constant.

Tot i això, per a poder iterar correctament, primerament cal realitzar una tasca de recol·lecció de dades (selecció del *dataset*) i obtenció de característiques inicials.

Concretament, caldrà realitzar una sèrie de passes ben definides per cada cicle. Noteu que, en finalitzar la darrera passa del cicle, es continua amb la primera de les descrites, iniciant un nou cicle:

1. Preparació de model i ajustament dels hiperparàmetres.
2. Avaluació i interpretació de resultats.
3. Estudi i comparativa entre models. Recerca de nova bibliografia en la direcció dels resultats.

Finalment, per concloure el projecte, caldrà realitzar la publicació de dades i redacció de la memòria del treball. Aquestes dades seran el resultat de diferents experiments, així com una anàlisi sobre el comportament final del sistema.

1.5 Planificació del projecte

La planificació temporal del projecte està subordinada a les diferents entregues parcials proposades a l'aula virtual. El procediment iteratiu descrit a l'apartat anterior es desglossa en les següents etapes:

Definició i planificació ***(03/03/19)*** Definir i establir una proposta general del projecte a desenvolupar. Es descriuen els objectius a alt nivell, així com les metodologies a aplicar per assolir-los.

Estat de l'art ***(24/03/19)*** Documentar i recopilar informació relativa a l'estat actual de les àrees de coneixement i tècniques necessàries per implementar el projecte. S'estudien tant les investigacions reconegudes com els models ja validats per la comunitat.

Disseny i implementació ***(19/05/19)*** Iterar durant diferents cicles de proposta de models i validació d'aquests. Cal una comparativa analítica de les diferents solucions per resoldre el problema i identificar les millors solucions.

Memòria ***(09/06/19)*** Redactar del document que detalla el procediment seguit per l'elaboració del projecte. Inclou la publicació final de dades i els diferents estudis analítics realitzats.

Presentació i defensa ***(16/06/19)*** Presentar i defensar davant tribunal el projecte realitzat. La defensa consta d'una presentació en format vídeo acompanyada d'una explicació del treball de l'alumne.

Capítol 2

Estat de l'art

La investigació acadèmica sobre detecció d'objectes a seqüències de vídeo es troba a un punt prou interessant. A més dels avanços continus en la investigació més teòrica [47], ja s'aplica en utilitats directes per la societat com l'ús de vehicles de conducció autònoma [20].

Actualment, cal descompondre el problema en dos grans blocs a tractar de forma independent: la detecció o reconeixement d'imatges i el seguiment o *tracking*. Ambdues àrees tenen un gran recorregut històric i actualment ja s'estan combinant en models complexos capaços d'analitzar l'evolució d'un element dins una seqüència d'imatges.

2.1 Detecció d'objectes

La detecció d'objectes en imatges és una tècnica que consisteix en la identificació de diferents elements en una imatge o fotografia. Aquesta identificació suposa, habitualment, trobar la localització i els llindars d'un objecte o element.

Si a més de detectar els objectes els hem de classificar, parlarem de reconeixement d'objectes en imatges. Aquesta tasca de reconeixement consisteix a detectar i identificar les diferents classes dels objectes presents a una imatge, com ara cotxes, persones o altres objectes quotidians.

Aquesta tècnica té un gran recorregut al llarg de la història de visió per computador. Es tracta d'una problemàtica que s'ha tractat de resoldre en múltiples ocasions mitjançant diferents propostes:

HOG (*Histogram of gradients*) Aplicant una *sliding window*, es genera un vector característic per cada fragment. Aquest es calcula a partir del gradient, intensitat i direcció dels píxels que el componen. L'estudi [13] presenta la solució basada en HoG amb un gran rendiment computacional que, amb combinació amb una *Support Vector Machines* (SVM) aconsegueix classificació d'objectes en temps real. Tot i això, el model presenta

certes mancances pel que fa a la detecció parcial d'elements i els objectes amb contorns suaus.

DPM (*Deformable Part Models*) El model proposat per [22] contempla la divisió dels elements en diferents parts. El model es basa en la idea dels HoG, però inclou la descomposició de l'objecte a detectar en varis sub-elements. Per això, el model identifica les parts i la localització de les mateixes que, en combinació, defineixen l'objecte a detectar. Tot i això, les estructures internes per identificar els objectes són relativament simples, i només permeten una jerarquia de dos nivells, pel qual no és possible detectar elements molt complexos.

R-CNN (*Region-based Convolutional Neural Networks*) Amb l'apogeu de les *Artificial Neural Networks* (ANN), sorgeix la temptació d'incorporar-ne l'ús al problema de la detecció d'objectes. Si bé la idea és encertada, el cost computacional és inabastable. Com a primera mesura [30] proposa dividir la imatge en potencials objectes, i només classifica aquests mitjançant ANN. Si bé el model va suposar una revolució pel que fa a la precisió dels resultats, el cost computacional no permet l'aplicació del sistema en temps real. Aquest sistema va marcar clarament una via d'estudi, sota la que es basen models reconeguts com *Spatial Pyramid Pooling* (SPP)[33], Fast R-CNN o Faster R-CNN.

Fast / Faster R-CNN Arran de la bondat de R-CNN sorgeixen diferents models com aquests que tracten de mantenir la qualitat en la predicció, alhora que acceleren el processament de cada fotograma. Si bé el principi és el mateix que en R-CNN, [29] i [73] proposen variacions al model com la unificació de les diferents regions sota un únic model o l'ús d'una ANN específica per la detecció de potencials segments d'imatge a classificar. Aquestes millores acceleren dràsticament el procés de classificació. Tant és així que la ràtio de rendiment respecte a la primera aproximació R-CNN respecte Faster R-CNN és de més de 250 (passant de 50 segons per imatge a 0,2 s).

YOLO (*You only Look Once*) A diferència dels models anteriors, [71] proposa un model que tracta tota la imatge com un únic element, independentment de cercar elements a diferents regions. Se subdivideix la imatge en una graella, on cada cel·la tracta de predir, d'entre les possibles classes, la que major coincidència comporti amb aquell segment concret. Tot seguit, es combinen els resultats de les cel·les per identificar les fronteres (*boundaries*) dels diferents elements detectats. Aquesta aproximació, tot i ser molt eficient per només haver d'avaluar cada segment un únic cop, té una sèrie de dificultats, com la detecció d'elements petits dins la imatge o una pitjor localització pel que fa a les coordenades dels elements detectats.

SSD (*Single Shot Detector*) Aquest darrer model, així com fa *You only Look once* (YOLO), segmenta la imatge en una graella per evitar la múltiple classificació d'una mateixa regió. [52] proposa el model cercant un equilibri encertat entre YOLO i la família R-CNN. Si bé no és tan eficient computacionalment com YOLO, la precisió augmenta aproximant-se als R-CNN.

Detectors emergents En l'actualitat segueixen sorgint diferents models bastats en les aproximacions anteriors, especialment en la cerca selectiva (com R-CNN) o *single shot* com YOLO. Entre els més prominents destaquen Mask R-CNN [32], RefineDet [93] o M2Det [94]. La direcció principal d'aquests models és avançar en un compromís entre qualitat de predicció i eficiència.

2.1.1 Mètriques d'avaluació

Aquests diferents models i aproximacions per a la detecció d'imatges competeixen entre ells per resoldre el problema de la millor manera possible. Tot plegat, el dubte que cal respondre ara és: què és resoldre el problema de detecció.

Per donar resposta a aquesta pregunta es recorre a diferents mètriques comunes entre les propostes per avaluar la bondat del model:

IoU (*Intersection over Union*) També coneguda com a índex Jaccard, és una de les mètriques més esteses per avaluar la localització i mida de les prediccions. El càlcul es realitza mitjançant la relació entre dues àrees: la caixa de predicció i la caixa de *ground truth*. La ràtio IoU és el resultant de dividir la intersecció de les àrees entre la unió de les mateixes.

$$\text{IoU} = \frac{\text{Àrea d'intersecció}}{\text{Àrea d'unió}}$$

Funció de classificació Els resultats de les funcions de classificació binària poden ser quatre: vertader positiu (TP), fals positiu (FP), vertader negatiu (TN) o fals negatiu (FN). La mètrica més bàsica sobre aquests possibles resultats és l'exactitud, que és la ràtio dels resultats vertaders entre el total. Arran d'aquests quatre possibles valors sorgeixen altres mètriques molt emprades com la precisió, l'exhaustivitat o F1 [66]. Aquestes, en lloc d'avaluar únicament la qualitat dels resultats, també mesuren quant concises són les prediccions entre els possibles valors.

mAP (*mean Average Precision*) La mètrica per antonomàsia per a la detecció d'objectes. Estableix relació entre mètriques clàssiques d'aprenentatge automàtic, com ara precisió

i exhaustivitat, amb altres pròpies de la visió per computador com IoU. Actualment, diferents *dataset* o conjunts de dades la fan servir per avaluar el rendiment del model a provar.

2.1.2 Conjunts de dades

Una vegada presentades les tècniques més rellevants trobades mitjançant la recerca, cal disposar de dades sobre les quals aquestes podrien executar-se.

Al llarg de les darreres dècades [25], diferents conjunts de dades i competicions s'han establert com a referents a l'hora d'avaluar les propostes que sorgeixen per resoldre la detecció d'objectes. Alguns dels més rellevants són COCO[51], PASCAL[18], ImageNet[16], Sun[90], INRIA[14], Caltech[17] o KITTI[26]. És clar que tots presenten característiques úniques i atributs que els fan diferenciar entre la resta. Això és interessant a l'hora de comparar models, ja que pot ser rellevant treballar en un context determinat, o per contra, cercar *datasets* generalistes.

2.2 Seguiment d'objectes

Per altra banda, el seguiment d'objectes és un concepte que aplica a les seqüències d'imatges o vídeos. Donada una imatge o fotograma amb un element detectat dins el mateix, un sistema de seguiment d'objectes s'encarrega d'estimar el moviment o trajectòria donat element al llarg dels fotogrames [92].

Històricament s'han presentat diferents aproximacions per aconseguir resoldre aquest problema, les quals s'engloben en les següents tècniques:

Fluxe òptic dens/espars: Es defineix, mitjançant diferents algorismes, un vector de moviment per cada un dels píxels o subconjunt dels mateixos. És una de les primeres aproximacions que ha quedat ja en desús.

Seguiment d'un únic objecte Aquesta categoria de *trackers* consisteix en, a partir d'un primer fotograma marcat amb una àrea a seguir. Tot i que es podria marcar manualment el segment a seguir, és habitual combinar aquesta tècnica amb un detector d'elements.

Seguiment de múltiples objectes Aquesta tècnica requereix detectors prou eficients. La idea principal consisteix a detectar els objectes a diferents fotogrames, i, mitjançant el seguidor, relacionar els objectes en el temps.

Entre les aproximacions descrites, és especialment rellevant centrar-nos en les dues darreres, que s'usen actualment el projectes d'avantguarda [91]. Aquestes dues tècniques acostumen a

referenciar-se per les seves sigles angleses *Visual Object Tracking* (VOT) i *Multiple Object Tracking* (MOT).

2.2.1 VOT (*Visual Object Tracking*)

El VOT consisteix en el seguiment d'un únic objecte al llarg d'una seqüència d'imatges. Durant les darreres dècades s'han produït diferents aproximacions per aconseguir realitzar seguiment visual d'objectes:

Una de les primeres incorporacions al seguiment d'objectes dins el camp de la visió per computador es va realitzar mitjançant el filtratge Kalman [74]. Aquest mètode existeix des de la dècada del 1960 amb aplicació directa sobre balística i guiatge de míssils [44]. Altres mètodes amb gran recorregut són Meanshift [12] i el seu derivat CamShift [5], que tracten de seguir l'objecte mitjançant la localització de la màxima densitat d'una funció. La principal mancança d'aquest mètode és la falta de robustesa respecte canvis bruscs de direcció.

Tot i això, en l'actualitat existeixen altres aproximacions per tractar aquest problema. Moltes d'aquestes es presenten al VOTChallenge [42], que congrega a gran part de la comunitat.

El VOTChallenge convoca un repte anual on es presenten diferents models punters pel que fa al seguiment visual d'un únic objecte. Durant les darreres convocatòries [41, 40, 39], els resultats han sigut prou interessants, ja competeixen models basats en diferents premisses relacionades amb models generatius/discriminants. Alguns models proposen *Convolutional Neural Networks* (CNN), altres SVM, o altres filtres de correlació discriminant. El punt rellevant d'aquests resultats és que és un problema tractat des de moltes perspectives diferents i cal experimentar amb múltiples alternatives.

Pel que fa a les mètriques d'avaluació als models VOT [87], i més concretament al VOTChallenge, destaca la *Expected Average Overlap* (EAO). La EAO tracta de combinar la mesura de bondats d'exactitud (*accuracy*) i robustesa del sistema.

Per avaluar els models proposats existeixen diferents conjunts de dades reconeguts i usats per la comunitat com TB50 [88], OTB [89], ALOV [79] o NUSPRO [49].

2.2.2 MOT (*Multiple Object Tracking*)

El MOT consisteix en el seguiment de múltiples objectes al llarg d'una seqüència d'imatges. Així com per VOT, és de gran importància el VOTChallenge, en aquest cas es troba el MOTChallenge [60].

A l'hora de dissenyar un sistema MOT, la complexitat augmenta respecte als models anteriors. Si bé un MOT pot ser interpretat simplement com una combinació o assemblat de varis VOT (una instància per element a seguir), existeixen algunes complicacions addicionals

relacionades directament amb la complexitat afegida respecte als VOT. Més específicament, en l'actualitat, la major feblesa dels MOT és l'excessiu volum de falsos negatius[47].

Pel que fa als sistemes MOT, la mètrica més estesa i general per avaluar-ne la bondat [2] és la *Multiple Object Tracking Accuracy* (MOTA). Aquesta mesura combina tres possibles fonts d'error: falsos positius, objectius perduts i intercanvis d'identitat.

En aquest cas, en relació als VOT, hom pot trobar menys *datasets* sobre els quals treballar. Tot i això, són importants els conjunts de dades de MOTChallenge [46, 61] o PathTrack [56].

2.3 Combinació de tècniques

Els algorismes de seguiment són, computacionalment parlant, molt més eficients que els de detecció. Tot i això, els de seguiment requereixen un estat inicial per identificar quin objecte seguir. A partir d'aquesta circumstància es pot establir una relació simbiòtica entre els dos conjunts de tècniques: l'ús de detectors i algorismes de detecció en un mateix sistema.

El principal benefici que aporta aquest assemblament és que els algorismes de detecció, més costosos, es poden executar en intervals de n fotogrames i que els de seguiment estableixin relacions entre els objectes detectats.

A més a més, un benefici que aconseguim d'aquesta combinació és la capacitat de, una vegada detectat un element, identificar com evoluciona aquest (aplicant seguiment) al llarg d'una seqüència. D'altra forma, només amb detectors, no podríem saber que hem identificat un element concret, ja que només podem saber la seva classe.

Tot plegat, cal dir que aquesta combinació és una pràctica ben estesa, que sense models de detecció, tot seguiment hauria de ser controlat de forma manual per identificar la regió d'interès.

Alguns dels sistemes que combinen detecció i seguiment més reconeguts són LSST17 [24], DS_v2 [61], amb grans qualificacions al MOTChallenge [46, 61], o també NOMT [8], JMC [85] o MDPNN16 [77], reconeguts per anàlisis comparatives de la comunitat acadèmica [47].

Tot plegat, en aquest punt del projecte, s'han descobert les diferents aproximacions per la detecció d'objectes. Aquestes es divideixen en aproximacions clàssiques de visió per computador, i les més recents, basades en *deep learning*. Pel que fa al seguiment d'objectes, existeixen també diferents aproximacions, d'entre les quals ens són rellevants les de seguiment de múltiples objectes (MOT).

Capítol 3

Descripció del mètode

Una vegada adquirit el coneixement sobre l'estat actual dels problemes de detecció i seguiment, es pot iniciar el procés de desenvolupament i experimentació sobre aquests.

L'objectiu principal d'aquesta etapa és la recapitulació i avaluació dels diferents models i tècniques actuals per la solució del problema, aplicant-lo a un conjunt de dades amb el qual s'està familiaritzat.

Tot plegat, el procés permetrà tant una millor comprensió de les tècniques i com l'ampliació amb nous coneixements segons es realitzen els experiments.

3.1 Procediment de disseny i implementació

Per al disseny i implementació del Treball Final s'ha seguit una metodologia iterativa, on els resultats dels prototipus de cada cicle han propiciat l'inici d'una següent passa en la investigació.

Com s'observa a la figura 3.1, aquests cicles s'han repartit en diferents blocs, executats de forma seqüencial i desenvolupats en detall al llarg del document:

1. Selecció de conjunt de dades o *dataset*.
2. Bloc de models de detecció d'objectes.
3. Bloc de models de seguiment d'objectes.

A més, cada bloc d'investigació sobre models, tant de detecció com de seguiment es descompon en les següents fases:

Investigació Recerca sobre les possibles vies obertes en l'àrea d'investigació actual.

Disseny Idealització de les tasques a realitzar, juntament amb l'objectiu de les mateixes.

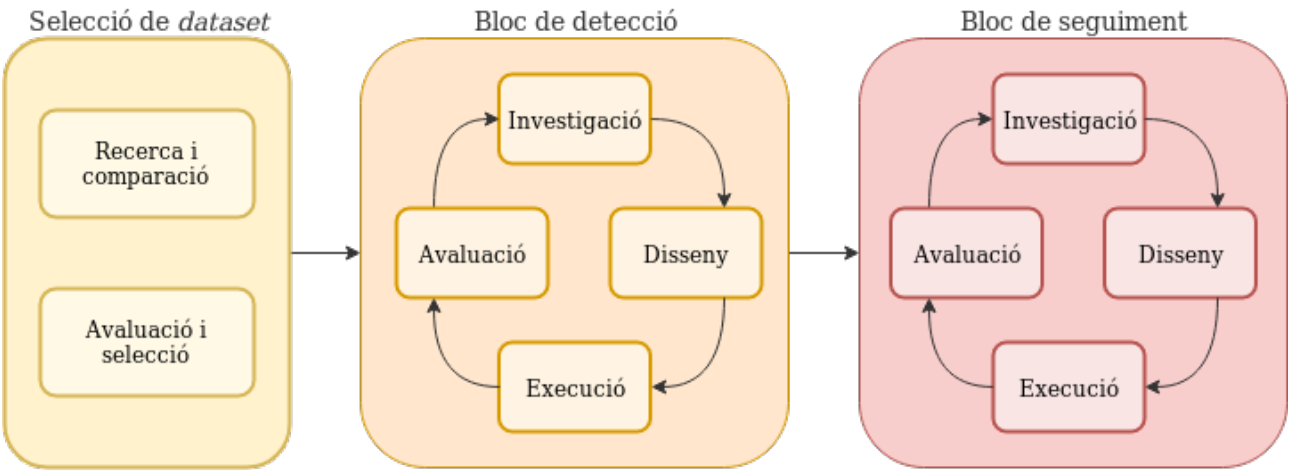

Figura 3.1: Procediment per a la realització del Treball Final.

Execució Implementació o ús de codi disponible per a la funcionalitat de l'experiment.

Avaluació Estudi de la bondat dels resultats de l'experiment, tant de forma qualitativa com quantitativa.

3.2 Obtenció de dades

Per al correcte desenvolupament i ús dels diferents mètodes i eines presentats, cal obtenir les dades de diferents escenes sobre les quals aplicar-hi els tècniques de detecció i seguiment. Abans de dissenyar experiments i implementar-los, ha calgut fer una recerca sobre els diferents conjunts de dades disponibles i la selecció d'un conjunt de dades o *dataset* sobre el qual realitzar diferents accions.

3.2.1 VAP Trimodal People Segmentation Dataset

La primera aproximació s'ha realitzat amb el dataset VAP Trimodal People Segmentation Dataset [65]. El punt més interessant d'aquest conjunt de dades és la presència de dimensions addicionals per descriure la realitat de l'escena. Així com la gran majoria de *datasets* representen només l'espai RGB, en aquest cas es disposa també de dues dimensions addicionals: la captura de profunditat i la tèrmica. La subsecció 3.2.1 il·lustra amb un clar exemple les diferències entre aquestes dimensions.

Tot i això, el conjunt de dades presenta algunes mancances que dificulten el desenvolupament àgil que requereix aquest treball. El principal d'aquests inconvenients és la falta d'un *ground truth* formal amb segregació d'elements. Si bé el conjunt de dades inclou màscares amb les

Figura 3.2: Mostra del mateix fotograma en les diferents dimensions: RGB, tèrmica i profunditat.

persones de les escenes, no ha sigut suficient com per a avaluar correctament els models de detecció, ja que és necessari diferenciar els elements entre si.

S'han intentat aplicar algunes tècniques automàtiques per dividir les persones de les màscares, però han sigut insuficients. A causa del gran volum de treball que suposaria l'etiquetatge manual del *ground truth*, s'ha decidit descartar el conjunt de dades.

3.2.2 MOTChallenge: MOT17

La MOTChallenge, ja presentada al secció 2.2, és una plataforma al voltant de la qual es pot trobar molta comunitat relacionada amb la visió per computador. La publicació de resultats de l'edició MOT17 [62], juntament amb una gran quantitat d'escenes ben etiquetades i definides varen fer d'aquest conjunt de dades la millor opció per a la realització dels diferents experiments del Treball Final.

El conjunt de dades disposa de set escenes d'entrenament i set més d'avaluació. Com es pot veure a la figura 3.3, aquestes escenes són diferents entre elles pel que fa a aspectes rellevants com la resolució, quantitat de persones, nivells d'oclusió, canvis d'il·luminació, etc.

A més de presentar un *ground truth* amb informació relativa a la posició dels elements i els seus identificadors únics (rellevant per la tasca de seguiment), també inclou deteccions realitzades sobre les escenes amb models preentrenats com *Deformable Part Models* (DPM), *Scale Dependent Pooling* (SDP) o *Faster RCNN* (F-RCNN).

Finalment, un darrer factor positiu per la selecció del conjunt de dades és el gran nombre de mètriques que recull, tant per la detecció com pel seguiment [80]. Aquestes es poden obtenir fàcilment, a partir de les prediccions dels nostres models, mitjançant el *kit* de desenvolupament oficial [48].

Figura 3.3: Mostra de diferents escenes del MOT17.

3.3 Detecció d'objectes

Una vegada seleccionada la col·lecció de dades a utilitzar, es poden començar a dissenyar i implementar diferents experiments. El primer bloc d'experiments consisteix en l'aplicació i comparativa de models de detecció d'objectes. Per aquesta tasca, s'han reduït les classes a detectar a només persones, ja que gran part del *dataset* MOT17 només contempla aquestes entitats.

3.3.1 Models de detecció

3.3.1.1 Deformable Parts Model (DPM)

Com bé s'ha presentat a la secció 2.1, les tècniques de *deep learning* s'han imposat a diferents camps de visió per computador durant els darrers anys. Tot i això, s'ha considerat oportú incloure a la comparativa altres mètodes clàssics que han demostrat el seu correcte funcionament.

Així doncs, el primer experiment consisteix en l'execució, sobre el *dataset*, d'un model DPM [23] per tal de detectar persones a les escenes d'entrenament. Recordem que es tracta d'un algorisme ben reconegut per la comunitat i amb múltiples implementacions obertes disponibles [70, 53].

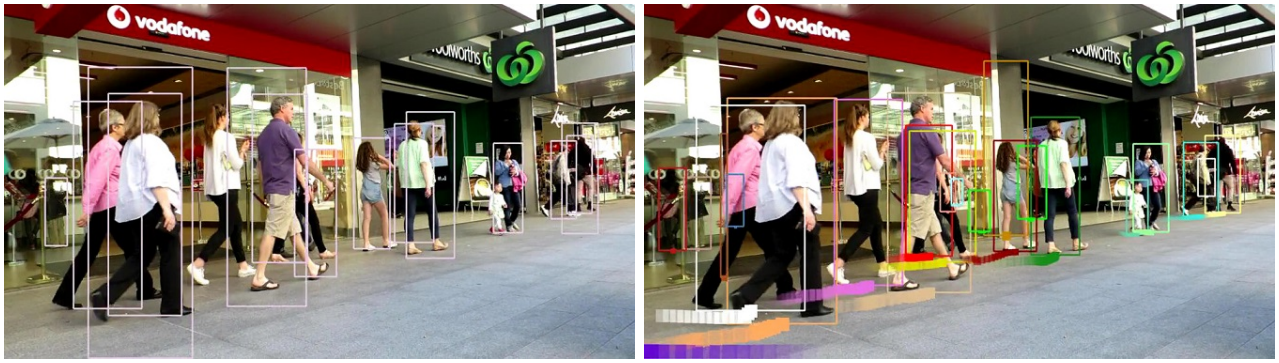

Figura 3.4: Deteccions DPM en relació al *ground truth*.

En aquest cas, no ha sigut necessari realitzar prediccions amb un model preentrenat, ja que la mateixa organització del MOTChallenge ofereix les prediccions realitzades per un model DPM sobre el *dataset* complet. A la figura 3.4 s'observen les deteccions del model i com es comparen amb el *ground truth*.

3.3.1.2 Scale Dependent Pooling (SDP)

Una vegada executada i analitzada l'aproximació del DPM, es percep la necessitat de començar a estudiar tècniques relacionades amb el *deep learning*. L'aproximació més senzilla és adaptar un classificador CNN a tècniques habituals de visió per computador com les finestres lliscants i el *pooling*. Durant la implementació d'aquest disseny, va sorgir un problema en relació a la gran majoria de CNN preentrades obertes [9] (com ara VGG16 [78], ResNet [34] o Inception [82]). Aquestes no contemplen cap classe per identificar persones, ja que s'entrenen amb el dataset ImageNet [76], així que es requeriria la implementació de tècniques com *transfer learning* [84]. Davant aquesta situació es va decidir descartar avançar per aquesta via, ja que s'allunya excessivament de l'enfocament del Treball Final.

Figura 3.5: Deteccions *SDP* en relació al *ground truth*.

Tot i això, com en el cas anterior, MOT17 inclou les prediccions d'un model basat en SDP.

Així doncs, es disposen de les dades necessàries per incloure a la comparativa de models de detecció aquesta tècnica. La figura 3.5 mostra diferències notables entre les àrees detectades i les esperades.

3.3.1.3 *Mask-RCNN*

Arran dels bons resultats descoberts amb tècniques de *deep learning*, es decideix seleccionar una aproximació puntera en l'estat de l'art actual. Com es va introduir en entregues anteriors, les R-CNN [82, 84, 28] resolen amb una bondat alta els problemes de detecció d'elements. Per això, d'entre les possibles variants [82, 27], es selecciona una de les més vigents: Mask-RCNN [31].

Figura 3.6: Deteccions Mask-RCNN en relació al *ground truth*.

Concretament, es fa servir la implementació oberta disponible a [58]. El sistema, a més de permetre un entrenament amb dades pròpies, inclou els pesos per un model preentranat amb el MS COCO [50]. Una vegada executat el model sobre el conjunt de dades s'han obtingut deteccions com les de la figura 3.6, amb bona qualitat pels elements en primer pla.

3.3.1.4 *YOLO v3*

El darrer dels experiments relacionats amb detecció es realitza amb el model YOLOv3 [72]. La principal diferència amb el R-CNN és que en aquest cas no se cerquen les regions rellevants abans de classificar, sinó que es divideix la imatge en una graella de cel·les dins les quals detectar objectes.

En aquest cas, la implementació concreta es pot trobar disponible a [68], basada en [72]. En aquesta darrera iteració sobre models de detecció, s'ha trobat un model capaç de realitzar prediccions, il·lustrades a la figura 3.7, prou precises amb un rendiment computacional acceptable per tasques de detecció i seguiment en temps real.

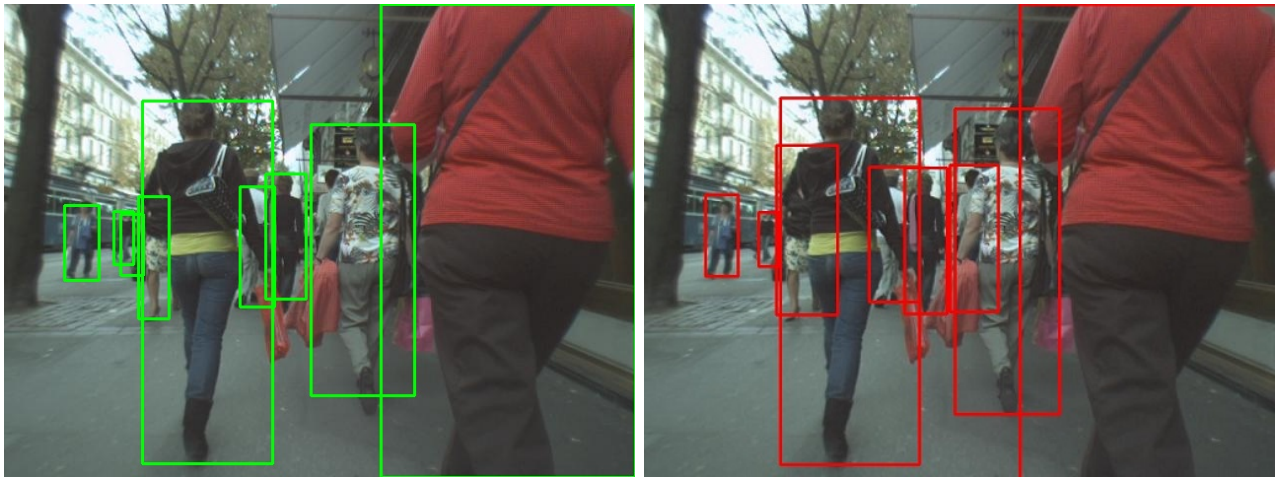

Figura 3.7: Deteccions YOLOv3 en relació al *ground truth*.

3.3.2 Mètriques de detecció

Per a l'avaluació dels diferents models a avaluar en els diferents experiments, es fan servir les següents mètriques, basades en les propostes del PASCALVOC [19]. Aquestes mètriques només apliquen a una única classe, ja que només es detecten persones. Això fa que no es calculin mètriques rellevants com mAP, ja que només tenen sentit per detectors amb múltiples classes:

Precisió Quantitat de prediccions positives que són correctes, respecte del total de prediccions.

Exhaustivitat Percentatge de casos positius detectats.

AP *(Average precission)* Precisió ponderada entre totes les mostres (fotogrames) del conjunt de dades (escena).

Deteccions per fotograma Mitjana de deteccions per fotograma de la seqüència.

Total de positius Nombre de deteccions realitzades pel model.

Matriu de confusió Recol·lecció de deteccions positives i negatives, tant veritables com falses.

Per a considerar una detecció com a positiva s'ha utilitzat la intersecció sobre unió (IoU) amb un llindar del 50%.

Totes les mètriques de detecció es calculen mitjançant la implementació oberta [69]. S'ha hagut de realitzar un procés d'*Extract, Transform and Load* (ETL) per tal de garantir consistència entre les diferents sortides de models i l'entrada del sistema calculador de mètriques.

3.4 Seguiment d'objectes

Una vegada desenvolupada la tasca de la detecció d'objectes, es decideix abordar el repte del seguiment d'objectes. Si bé tenen una forta relació entre ells, els dos problemes calen ser resolts amb tècniques ben diferenciades.

Tot i que no és una condició indispensable per realitzar seguiment, en aquest cas s'han utilitzat les deteccions obtingudes com a resultat dels experiments anteriors. Això ens permet no haver d'identificar manualment l'àrea d'interès per iniciar el seguiment.

3.4.1 Models de seguiment

3.4.1.1 CamShift

D'entre els diferents models introduïts a l'estat de l'art, el mètode *Continuously Adaptive Mean Shift* (CamShift), és un dels clàssics amb bona acceptació per la comunitat. Així doncs, amb un punt de partida sòlid, s'inicia el primer experiment del bloc de seguiment.

Figura 3.8: Seguiment en CamShift pels fotogrames: 1, 10, 50 i 100.

Si bé és cert que en alguns casos i escenes el seguiment és acceptable, en general és insuficient. Com podem observar a la figura 3.8, algunes de les deteccions s'expandeixen fins a perdre completament la referència original.

Aquesta anomalia succeeix quan part de la regió d'interès per la instància de seguiment conté fons, o és fàcilment confusible amb el mateix. El motiu d'aquest comportament erroni és que el model genera màscares basades en la tonalitat (*hue*) de la regió d'interès. Si la tonalitat no és prou identificable (perquè la mitjana de píxels contempla regions d'interès i fons), aquesta serà de poc valor.

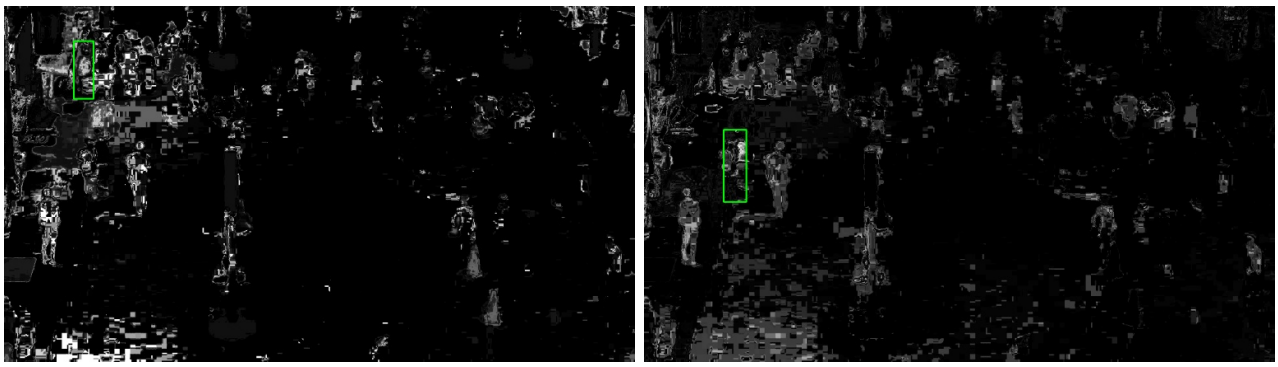

Figura 3.9: Mostres amb alt nivell de renou al voltant de la ROI.

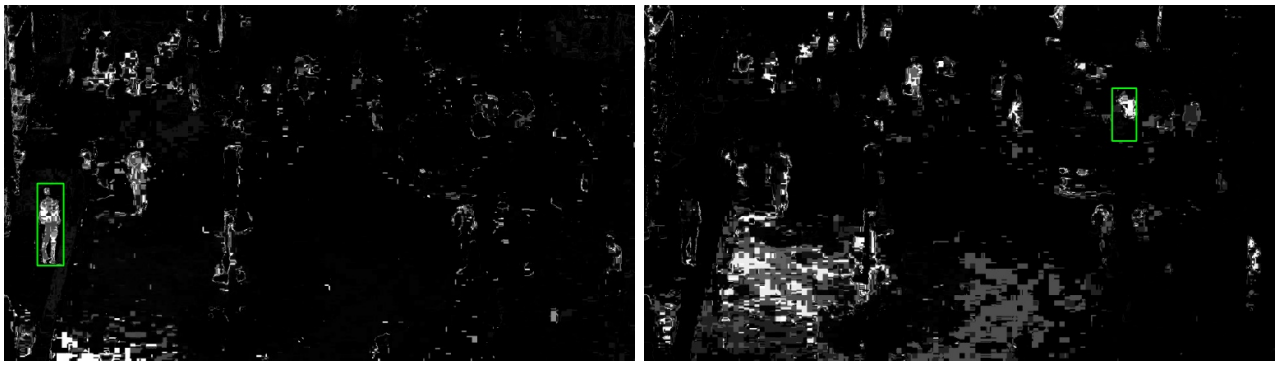

Figura 3.10: Mostres amb baix nivell de renou al voltant de la ROI.

Per avaluar aquesta teoria es mostren algunes de les màscares conflictives i la seva màscara per l'algorisme de seguiment. A la figura 3.9 es pot comprovar una gran semblança entre els elements rellevants i els seus voltants, mentre que a la figura 3.10 regions d'interès que no es confonen amb el fons que les envolta.

A causa de la baixa qualitat dels resultats, s'opta per descartar la via del mètode CamShift i s'inicia un altre experiment sense relació directa amb aquest sistema.

3.4.1.2 Filtre de correlació

Una altra tècnica present en els sistemes de seguiment són els filtres de correlació. Si bé no és una tècnica innovadora, la proposta [15] els fa servir per estimar el seguiment dividint el problema en dues parts independents: la variació d'escala i el moviment.

Figura 3.11: Seguiment en filtre de correlació pels fotogrames: 1, 10, 50, 100, 500 i 1000.

Així doncs, el mètode descrit es troba implementat a la llibreria Dlib [38], i proposa una API molt senzilla d'usar per experimentar ràpidament. En aquest cas, sota les mateixes condicions que l'experiment anterior, s'observen unes prediccions molt més robustes.

A la figura 3.11 es pot observar una robustesa millor que en l'experiment anterior, però encara hi ha una sèrie de problemes presents:

- Es perd el seguiment en produir-se interseccions entre elements o oclusions dels mateixos.
- Els objectes a seguir que desapareixen de l'escena no són identificats. Com a conseqüència directa, el model de seguiment roman a l'espera de canvis a una regió on no s'hi troba cap element rellevant.
- Donat que es parteix d'un conjunt de deteccions inicials, és impossible seguir elements que s'incorporen a l'escena en un instant que no sigui el primer fotograma.

3.4.1.3 SORT

Per resoldre dues de les tres mancances detectades a les aproximacions anteriors, cal detectar objectes en més instants que no només el fotograma inicial. Si bé és habitual combinar detectors i seguidors en un únic sistema, en aquest experiment s'ha decidit implementar una aproximació radicalment diferent. En lloc de realitzar seguiments basats en les dimensions de color (RGB) i la seva evolució, mitjançant el mètode *Simple Online and Realtime Tracking* (SORT) [3], ara es realitza una tècnica anomenada seguiment mitjançant detecció.

El seguiment mitjançant detecció (*tracking by detection*) [21] consisteix a detectar els elements rellevants a tots i cada un dels fotogrames amb un detector i, posteriorment, establim relacions entre les deteccions del fotograma actual i el predecessor.

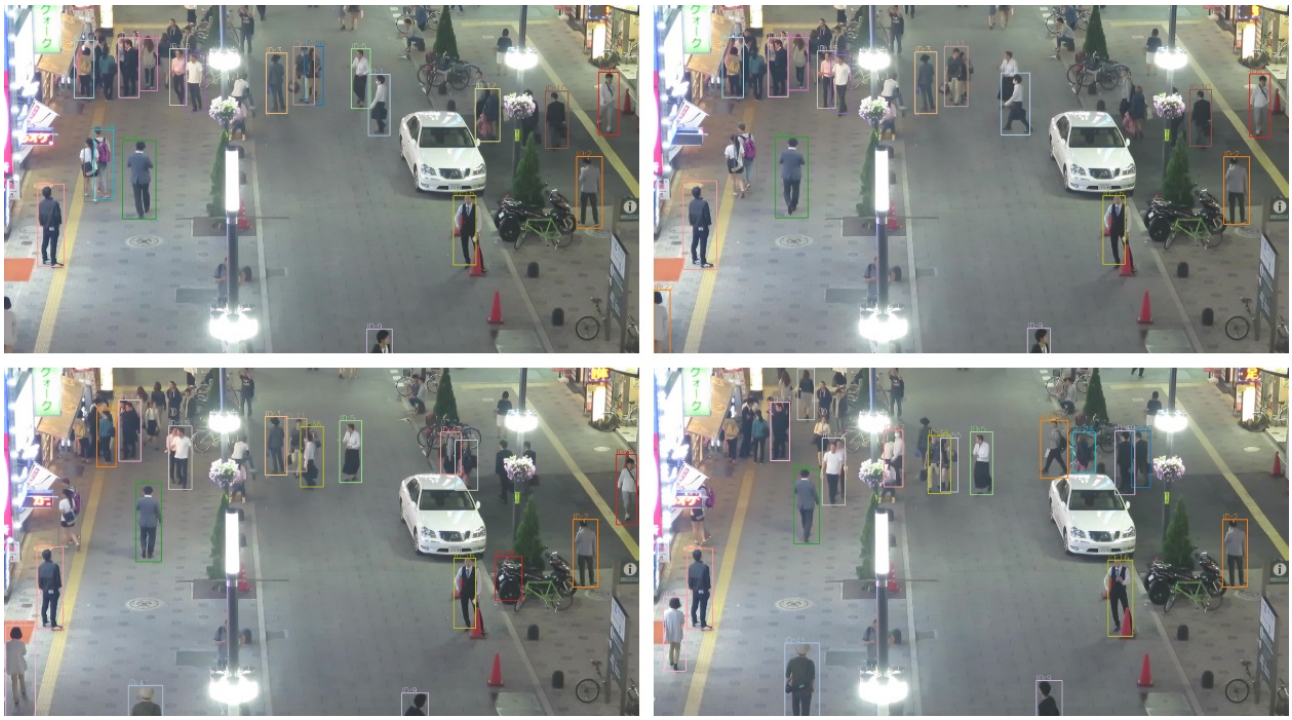

Figura 3.12: Seguiment en SORT pels fotogrames: 1, 10, 50 i 100.

Concretament, en aquest cas, s'han seleccionat les deteccions obtingudes amb el model YOLOv3 per relacionar-les entre si i definir seguiment entre objectes únics. Per a establir les relacions entre deteccions de fotogrames es fa servir l'algorisme hongarès [43], que tracta de cercar un mínim global pel que fa a les distàncies entre les deteccions de dos fotogrames consecutius. A la figura 3.12 s'exemplifica l'evolució dels objectes rellevants, que mantenen identificador i color, mentre el sistema d'assignació n'estableix la relació per cada instant.

3.4.1.4 SORT i filtre de correlació

L'experiment de seguiment mitjançant detecció presenta bons resultats, però no resulta, computacionalment, possible treballar en temps real amb alguns models de detecció. Com a conseqüència, una pràctica ben estesa és la combinació entre models de detecció i models de seguiment sota un únic sistema.

En aquest darrer experiment s'investiga la possibilitat de combinar els filtres de correlació amb l'assignació SORT, que han propiciat bons resultats per separat. En aquest sistema es basa en el concepte senzill d'establir una freqüència determinada per detectar objectes rellevants, mentre que durant la resta de fotogrames es fa servir l'algorisme de seguiment que pertoqui.

Figura 3.13: Seguiment en SORT + filtre de correlació pels fotogrames: 1, 10, 50 i 100.

Per aquesta implementació [96] s'han fet servir les deteccions de YOLOv3 i el detector fa servir, de forma aleatoritzada amb una probabilitat de 0,4. Això implica que el detector només actua al 40% dels fotogrames, alliberant l'ocupació de recursos de la màquina on s'executa el sistema. A la figura 3.13 s'observen diferents instants durant el seguiment del model, el primer dels quals es basa en la detecció per delimitar regions d'interès.

3.4.2 Mètriques de seguiment

Per a l'avaluació dels diferents models a avaluar en els diferents experiments, es fan servir les següents mètriques, basades en les propostes del MOTChallenge [96].

MOTA (*Multi-Object Tracking Accuracy*) Mètrica general per avaluar la bondat d'un sistema de seguiment múltiple d'objectes. Representa tres possibles errors: el nombre de deteccions perdudes, el nombre de falsos positius i el nombre d'identificacions incorrectes (mal assignades a l'objecte que corresponen).

IDF1 Mesura combinatòria de precisió i exhaustivitat. En aquest cas, s'agrupa per ID d'objecte seguit.

Camins principalment seguits Nombre de camins seguits durant, al manco, un 80% del seu recorregut segons el *ground truth*.

Camins principalment perduts Nombre de camins seguits durant, com a màxim, un 20% del seu recorregut segons el *ground truth*.

Matriu de confusió Recol·lecció de deteccions positives i negatives, tant veritables com falses.

Intercanvi d'ID Nombre d'ocasions en què s'ha intercanviat, erròniament, l'identificador d'un mateix objecte.

Fragmentació de camins Nombre d'ocasions en què un trajectòria es veu fragmentada en algunes prediccions diferents (per exemple, perduda de *tracking*).

Aquestes mètriques es calculen mitjançant la implementació oberta [7]. S'ha hagut de realitzar un procés d'ETL per tal de garantir consistència entre les diferents sortides de models i l'entrada del sistema calculador de mètriques.

Capítol 4

Experiments i avaluació

Una vegada presentades i implementades les diferents tècniques i eines requerides per a la realització de l'estudi, cal analitzar-ne el comportament. Per assolir aquesta tasca, s'han dut a terme una sèrie d'experiments, tots fent servir la base de dades del MOT17.

Concretament, es volen validar les diferències i hipòtesi generades en el capítol anterior, i arribar a una conclusió sobre els avantatges i inconvenients de les diferents aproximacions disponibles en l'actualitat per resoldre els dos problemes tractats en aquest Treball Final.

4.1 Escenes dels experiments

Per avaluar els models proposats, tant de detecció com de seguiment, es faran servir les escenes del conjunt d'entrenament del MOT17. Aquest conjunt de dades es compon per un total de set escenes amb diferents característiques i problemàtiques habituals a l'hora de treballar en visió per computador.

Nom	FPS	Resolució	Fotogrames	Trajectòries	Deteccions	Densitat
MOT17-02	30	1920x1080	600	62	18581	31.0
MOT17-04	30	1920x1080	1050	83	47557	45.3
MOT17-05	14	640x480	837	133	6917	8.3
MOT17-09	30	1920x1080	525	26	5325	10.1
MOT17-10	30	1920x1080	654	57	12839	19.6
MOT17-11	30	1920x1080	900	75	9436	10.5
MOT17-13	25	1920x1080	750	110	11642	15.5

Taula 4.1: Descripció tècnica dels atributs de les escenes

A la taula 4.1 es disposen les característiques tècniques de les diferents escenes amb aspectes tan rellevants com la resolució de les imatges, els fotogrames per segon o la densitat mitjana

d'elements per fotograma. Tot i això, considero que és encara més rellevant i necessària una descripció qualitativa de les escenes i les característiques rellevants per als problemes a resoldre.

MOT17-02 S'observen diferents persones passejant per una plaça. La càmera es troba estàtica des d'una posició frontal. No hi ha grans canvis de lluminositat i el principal potencial problema és la superposició de vianants que s'entrecreuen en diferents fotogrames.

MOT17-04 Gran quantitat de persones es mouen per un carrer al vespre. La il·luminació és artificial però no homogènia, el qual pot canviar la il·luminació d'un mateix objecte detectat. Càmera estàtica des d'una perspectiva zenital.

MOT17-05 Una càmera mòbil avança per un carrer amb vianants que entren i surten d'escena. La posició de la càmera respecte dels vianants és frontal i no es troben grans canvis de lluminositat.

MOT17-09 S'observa un carrer sense vehicles amb diferents comerços, amb persones als interiors, a més de vianants fora dels mateixos. La càmera es troba a una posició frontal respecte els vianants, a més d'estar estàtica. Un potencial problema per a la correcta detecció són els reflexos de diferents vidrieres.

MOT17-10 S'observa un carrer, de vespre, amb vianants. La il·luminació no es manté homogènia. En aquest cas, la càmera és mòbil i sembla portada per una persona, el qual genera oscil·lacions verticals entre fotogrames. Aquesta particularitat també provoca que alguns fotogrames no estiguin ben enfocats, generant contorns difusos pels vianants.

MOT17-11 S'observa l'interior d'uns grans magatzems. La llum és artificial però homogènia. La càmera, en moviment, enregistra de forma frontal les persones que passegen, sense oscil·lacions brusques. Tot i això, tan diferents vidrieres com el sòl polit reflecteixen les siluetes de les persones.

MOT17-13 Escena d'un carrer amb trànsit de vianants i de vehicles, sota llum diürna. La càmera enregistra l'escena des d'un vehicle en circulació, el qual genera oscil·lacions tant verticals com horitzontals. A més a més, es produeixen canvis complets de plans, el qual fa que dins la mateixa escena les condicions siguin variants.

Com es pot comprovar, el conjunt de dades es conforma de diferents escenes amb característiques ben definides i diferenciades entre elles. Aquestes serviran per avaluar les virtuts i mancances dels models, tant de detecció com de seguiment. A més, també són útils per contrastar si els models són semblants entre ells pel que fa a la resposta d'una mateixa entrada (cada una de les escenes).

4.2 Experiments de detecció

Com s'ha introduït al secció 3.3, disposem de quatre models capaços de detectar persones a una imatge. Per a la realització dels experiments de detecció, s'han utilitzat tots quatre models per trobar persones a les escenes on s'esperen vianants.

Aquest estudi es divideix en dos blocs: un primer on es realitza un estudi numèric, i el segon on es realitza una interpretació qualitativa dels motius i relacions entre comportaments de models.

Pel que fa a l'avaluació de mètriques, a la taula 4.2 s'observa el comportament mitjà de cada un dels models sobre tot el conjunt de dades. D'entre les mètriques disponibles destaquen l'*Average Precision* (AP) i l'*F1*, que són mètriques ponderades que ens indiquen, de manera general la bondat del model.

Model	F1	AP	Precisió	Exhausitivitat	Det. veritables	Det. falses
DPM	0.295	0.294	0.883	0.196	51351	28439
SDP	**0.358**	**0.454**	**0.994**	**0.238**	**79842**	**2945**
Mask-RCNN	0.188	0.193	0.892	0.118	22529	6589
YOLOv3	0.337	0.354	0.883	0.234	57244	22134

Taula 4.2: Mètriques mitjanes de cada model per totes les escenes.

Es pot observar com el rang del valor F1 es situa entre [0.188, 0.358]. Pot semblar que la bondat dels models és baixa, però cal contextualitzar la problemàtica que s'estudia. Els 15 millors resultats de la competició presenten una AP d'entre [0,62; 0,89]. Cal entendre que aquests models han sigut entrenats i ajustats específicament per resoldre aquest repte. Per contra, els models presentats per a la realització de l'experiment són models generals sense ajustament d'hiperparàmetres ni tècniques *fine-tuning*. Així doncs, llevat de la qualitat absoluta de les prediccions dels models, es procedeix a una comparativa entre els mateixos.

D'entre les diferents mètriques de la taula 4.2, resulten especialment destacables dos punts:

- L'aproximació de DPM, tot i no ser tan actual com les basades en CNN, es manté vigent amb una bondat aproximada a models recents com YOLO. Això ens indica que, al contrari del que pugui semblar per l'actualitat divulgativa, no cal aplicar *machine learning* per resoldre tots els problemes, sinó que algunes aproximacions específiques anteriors encara són útils.

- El baix rendiment del model Mask-RCNN és destacable. Segons diferents estudis [73, 95], és un dels models més punters pel que fa a la detecció d'objectes. Tal vegada seria encertat realitzar experiments addicionals per comprovar que no es tracta d'una problemàtica

d'ajustament d'hiperparàmetres. Una altra opció és que les regions d'interès no es trobin acuradament a causa de la variació de dimensions entre els objectes a detectar, els vianants.

A la figura 4.1 es veu clarament que les característiques de les escenes afecten a la qualitat de les prediccions. S'observen pics a les escenes 5 i 11, mentre que les escenes 2 i 4 presenten pitjors resultats que la mitjana. Un fet prou interessant és que les dues mètriques presenten un comportament similar, el qual és un bon indicador de que són mètriques ben generalitzades i harmonitzades respecte mètriques més concretes o de baix nivell.

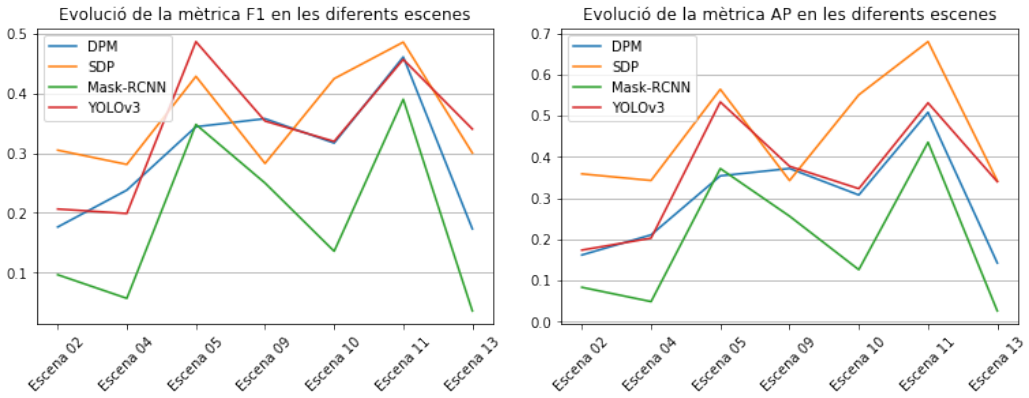

Figura 4.1: Comparativa entre escenes de les principals mètriques de detecció.

A més de les semblances de comportament entre escenes, a la figura 4.1 també s'observa una jerarquia qualitativa entre els models: Mask-RCNN presenta els pitjors resultats a totes les escenes, mentre que DPM és més uniforme entre les diferents entrades. Per altra banda, els models amb millor rendiment són SDP i YOLOv3.

Per analitzar en detall la resposta dels models a les característiques de les diferents escenes, és prou útil descompondre les gràfiques en una única per model, on poder-ne observar el comportament per escena.

A la figura 4.2 es poden estudiar les bondats relatives per escena de cada model. Observem com el principal problema als models és una baixa exhaustivitat, ja que les prediccions són bones (totes per sobre del 0,8). Això implica que molts dels objectes presents al *ground truth* no són reconeguts pels detectors proposats. En aquest punt, l'anàlisi dels resultats de detecció es divideix en dues vies: l'estudi de les diferències entre escenes per tal de descobrir possibles característiques beneficioses pels detectors i l'estudi de la baixa exhaustivitat de tots els models.

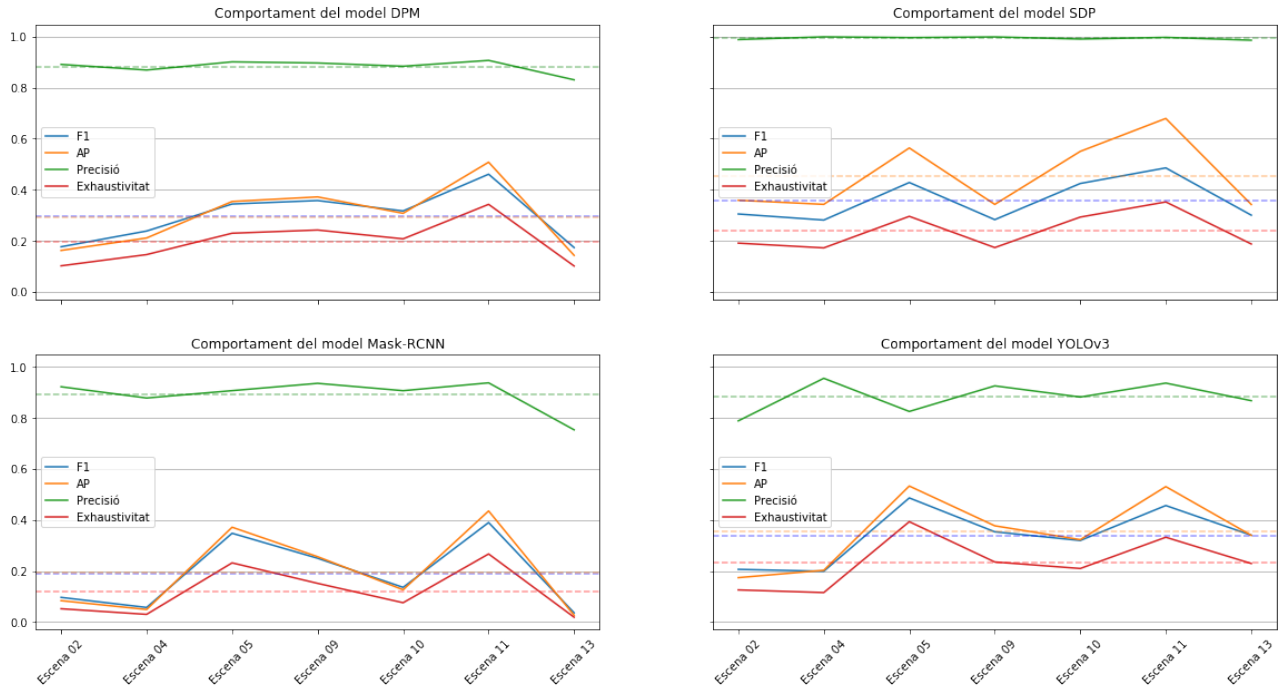

Figura 4.2: Comportament dels models segons les escenes processades.

4.2.1 Propietats de les escenes

Com bé s'ha explicat a la secció 4.2, els models presenten bondats altes per algunes escenes, i bondats molt baixes per altres escenes. Concretament observem que les escenes 05 i 11 estan per sobre de la mitjana de la mètrica F1 a tots els models, mentre que les 02, 04 i 13 estan sempre per sota de la mitjana.

A la figura 4.3 es mostren fotogrames de les dues escenes amb major qualitat de detecció entre tots els models. Si n'observem les característiques més rellevants, veiem que les càmeres es troben en moviment i que les persones que hi apareixen ocupen un espai vertical bastant elevat. Això vol dir que, com que les persones estan properes a l'objectiu de la càmera l'espai que ocupen i el nivell de detall són majors. A més, la il·luminació en ambdues escenes és bastant homogènia i estable. Per altra banda, un aspecte que a priori es podria considerar negatiu és que les persones de les escenes es creuen entre elles, ocultant en algunes ocasions vianants més allunyats rere els que es troben propers.

A la figura 4.4 es mostren fotogrames de les escenes 02, 04 i 13. Aquestes són les que pitjors resultats han generat amb els diferents models. Les característiques més visibles són diferents entre elles: càmeres mòbils i estàtiques, diferents tipus d'il·luminacions, etc. La primera escena, MOT17-02 és prou semblant a les presentades a la figura 4.3, ja que les persones en primer pla interactuen i tenen una mida semblant a altres escenes amb bones deteccions. Tot i això,

Figura 4.3: Fotogrames de les escenes 05 i 11.

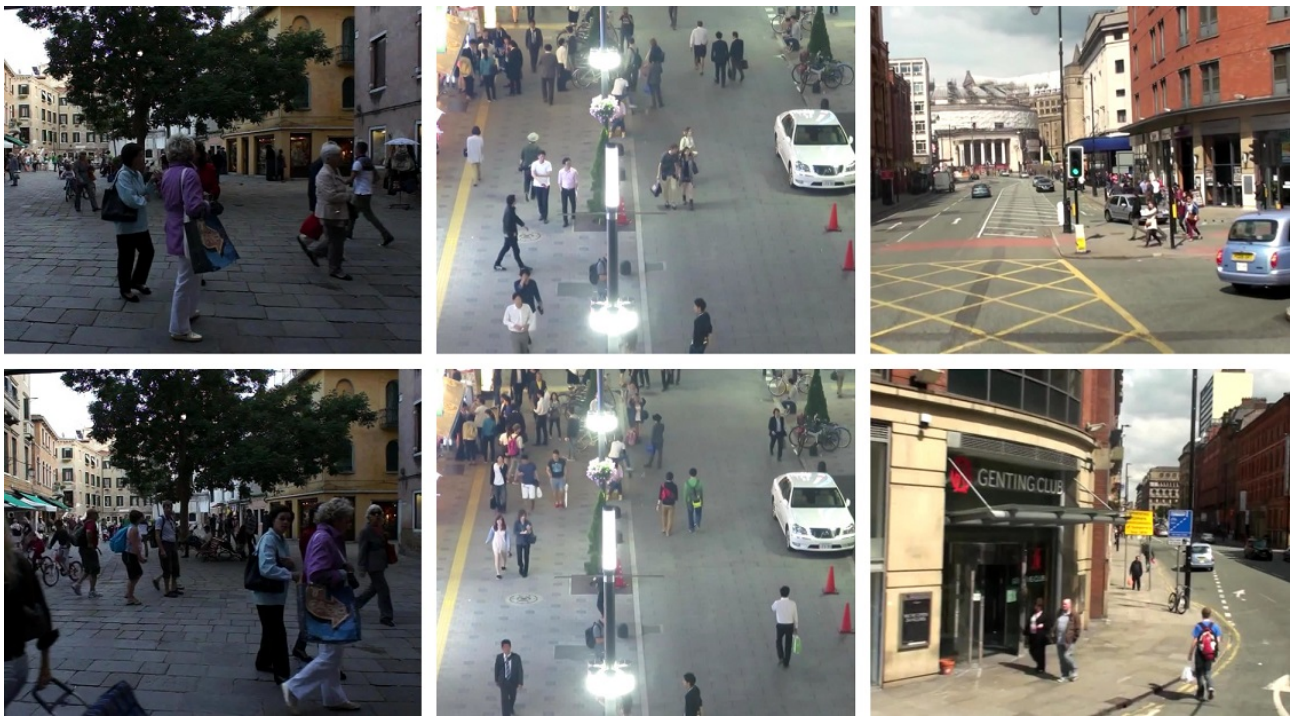

Figura 4.4: Fotogrames de les escenes 02, 04 i 13.

aquesta escena no presenta una il·luminació tan bona com les anteriors, i alguns colors de figures rellevants es poden confondre amb el fons que les envolten.

El que sí que tenen en comú, i és un punt diferencial respecte les escenes anteriors: les mides de les persones són altament variables. Podem trobar moltes persones alhora amb diferents mides, el qual pot ser un potencial problema per alguns models, especialment si la gestió d'escales no es duu a terme correctament.

Tot plegat, les diferències entre les escenes amb millors i pitjors deteccions semblen poder-se resumir en tres punts principals:

Mida dels objectes a detectar. Les escenes amb pitjors mètriques presenten vianants amb mesures fluctuants, però en totes elles apareixen persones de mida petita en relació al fotograma complet.

Il·luminació pobre o inconsistent. En les escenes de baixa qualitat la il·luminació no és consistent en tots els casos, o no permet diferenciar les persones amb gran contrast de saturació. Pel que fa a les escenes bones, la il·luminació és intensa i els colors ressalten respecte al fons.

Congregació de multituds. En el cas de les escenes presentades a la figura 4.4, de baixa qualitat, trobem en molts de casos grups de persones que s'acumulen en diferents punts. Una possible explicació de per què aquesta característica afecta negativament a la detecció dels models és que les regions tan compactades presenten més elements que els models poden detectar en aquella zona concreta.

4.2.2 Baixa exhaustivitat

Per analitzar la baixa exhaustivitat, primerament cal analitzar el *ground truth*. Si bé el *dataset* està reconegut per la comunitat, tal vegada les escenes inclouen objectes de molt difícil detecció, o que oclusionen entre ells, i per tant el detector no pot identificar en certs fotogrames.

A la figura 4.5 es veuen ben definits els contorns de les persones. S'observen alguns solapaments entre caixes, principalment deguts a objectes superposats en el fotograma. Es veu com les reflexions sota les persones no estan identificades com a objectes veritables, com és d'esperar en el *ground truth*.

A la figura 4.6 s'observen, novament caixes amb persones ben identificades a l'interior. Tot i això, no només es tornen a produir solapaments, sinó que també es detecten com a vàlids els reflexos als cristalls. Si bé és una decisió discutible, el fet que aquests elements siguin part del *ground truth* dificulta molt la bona puntuació dels models, ja que són estímuls gairebé imperceptibles fins i tot per humans.

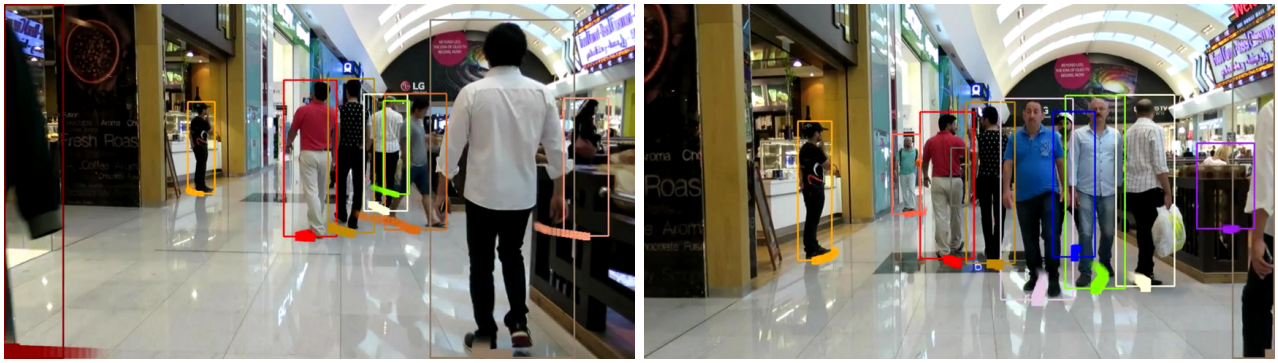

Figura 4.5: Fotogrames del *ground truth* de l'escena MOT17-11.

Figura 4.6: Fotogrames del *ground truth* de l'escena MOT17-09.

A la figura 4.7, que mostra l'escena 13, s'observen una sèrie de caixes que no es corresponen amb l'objectiu de l'estudi. Si bé gran quantitat de vianants són etiquetats, també apareixen vehicles, senyals de trànsit o fins i tot el pal d'un semàfor. Totes aquestes deteccions potencials no són contemplades pels models presentats.

A la figura 4.8 es presenta un exemple extrem d'escena amb més elements dels esperats. Com es pot observar no només s'identifiquen grups amb un nombre excessiu de persones que col·lisionen (cantó superior esquerre) sinó que es detecten una gran quantitat d'objectes fora de l'estudi dels nostres models, com ara vehicles, fanals, senyals o cons.

Tot plegat es descobreix que la baixa exhaustivitat pot ser causada per les caixes no relacionades amb persones que es troben al *ground truth* d'algunes escenes. Totes aquestes deteccions fan que l'anàlisi de mètriques quantitatives dels models proposats, que només detecten persones, sigui poc acurada. Així doncs, donat que no és possible, amb els mitjans actuals, avaluar els models únicament sobre el *ground truth* el més interessant és comparar-los entre ells.

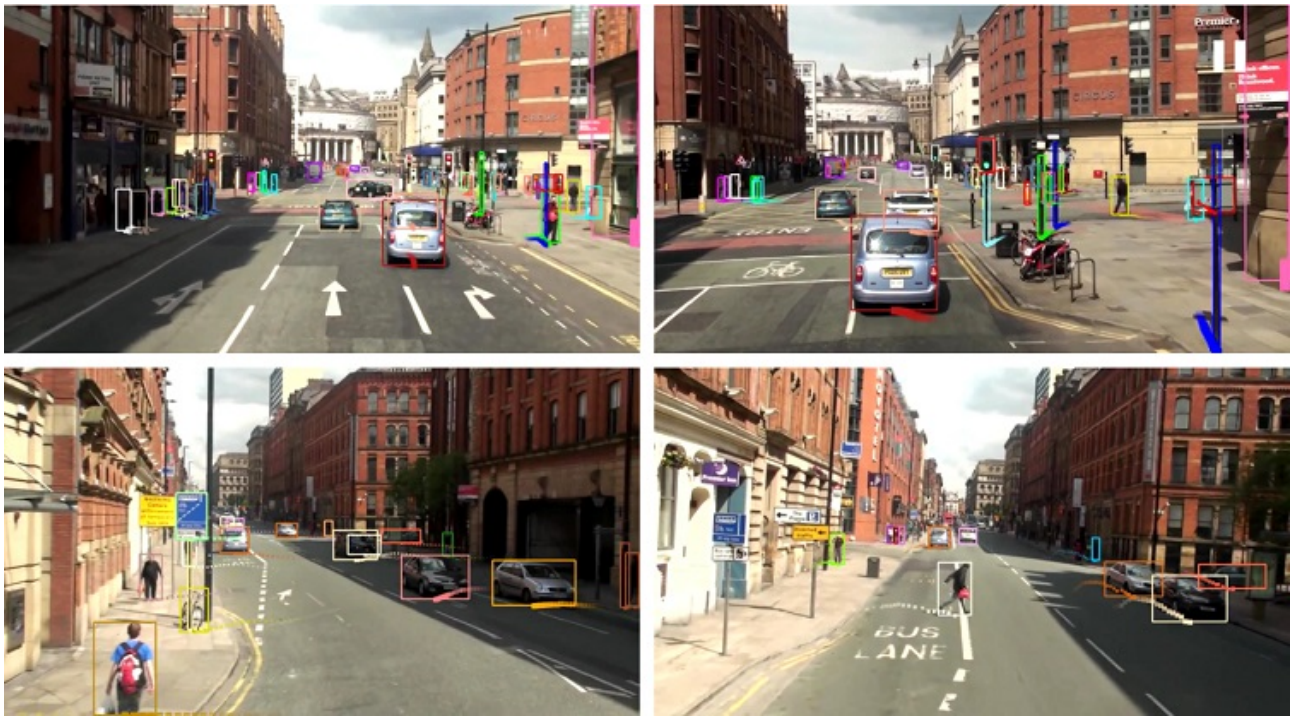

Figura 4.7: Fotogrames del *ground truth* de l'escena MOT17-13.

Figura 4.8: Fotogrames del *ground truth* de l'escena MOT17-04.

4.2.3 Conclusions dels experiments de detecció

Una vegada realitzats els diferents experiments relatius a la detecció d'objectes i analitzats els resultats, es consoliden diferents idees relatives a aquestes tècniques i quins resultats pot

esperar hom en fer-les servir.

La primera de totes és que cal cercar l'aproximació més indicada pel problema concret que es vulgui resoldre. Com s'ha comprovat en aquests experiments, alguns models com SDP encara són vigents i poden ser més efectius que aproximacions modernes de *deep learning*.

El segon aprenentatge és la importància de la natura de l'escena a l'hora de realitzar les deteccions. Si hom coneix el context de les dades sobre les quals el sistema treballarà, pot adaptar amb major precisió tant les configuracions com les arquitectures dels models.

Finalment, cal no descuidar esforços necessaris en procediments manuals, com l'etiquetatge del *ground truth*, ja que acostumen a ser crítics a l'hora d'avaluar models d'aprenentatge automàtic.

4.3 Experiments de seguiment

Una vegada compresos els resultats dels models de detecció, es procedeix a estudiar els diferents experiments de seguiment d'objectes. Si bé, com s'ha introduït a la descripció del mètode, no cal lligar el seguiment a la detecció, en aquest cas s'han seleccionat els dos millors models de detecció (SDP i YOLOv3) per aplicar-ne seguiment a les deteccions que han oferit.

4.3.1 Models vàlids per l'estudi

Com ja s'ha introduït a la secció 3.4, els models sense assignació de relacions entre deteccions han resultat molt pobres i només resolen el problema de forma parcial. Recordem que aquests són incapaços de tractar objectes que entren o surten en escena en diferents instants. Aquestes mancances exclouen l'aproximació CamShift i de filtres de correlació (sense SORT) de la següent anàlisi.

Així doncs, es disposa de dos models de seguiment: el SORT, basat en seguiment mitjançant deteccions i la combinació entre SORT i filtres de correlació. Ambdós presenten mancances i virtuts que caldria considerar a l'hora de portat un sistema a producció:

SORT Com aspecte positiu, el cost computacional és molt baix, ja que no necessita realitzar cap tipus de tractament sobre les dades originals, les imatges. En aquest cas únicament es treballa amb regions de detecció (proveïdes per un sistema de detecció), el qual suposa alhora una contrapartida: la bondat del sistema recau principalment sobre el detector, deixant la tasca de seguiment en un segon pla.

SORT amb filtre de correlació Aquest sistema híbrid permet realitzar de forma intel·ligent dues tasques alhora: combinar les deteccions entre elles i realitzar seguiment visual dels

objectes quan hom no disposa de deteccions en un fotograma. Si bé pot semblar que el potencial d'aquest model és inferior al SORT, és rellevant entendre que la velocitat de processament és considerablement superior, ja que no cal realitzar deteccions en cada fotograma. A més, com ja s'ha comprovat a la descripció del mètode i a [4], els algorismes de seguiment visual tenen un rendiment satisfactori a l'hora de resoldre el problema descrit.

4.3.2 Rendiment de les propostes

Deteccions	Model seguiment	MOTA	idF1	Precisió	Exhaustivitat	#Inter. ID
SDP	SORT	**0.555**	**0.532**	**0.944**	0.598	908
SDP	combined	0.529	0.531	0.881	**0.617**	820
YOLO	SORT	0.347	0.377	0.824	0.448	1011
YOLO	combined	0.294	0.389	0.743	0.457	**773**

Taula 4.3: Índexs i mètriques de bondat per la tasca de seguiment d'objectes.

A la taula 4.3 es mostren les principals mètriques per a l'avaluació de bondat dels models. En consonància amb la hipòtesi inicial, s'observa com existeix una relació directa entre la qualitat de les deteccions i els models de seguiment que les fan servir. En aquest punt és prou interessant comparar els models de detecció que es basen en les mateixes deteccions, més que aprofundir entre les diferències generades per deteccions diferents.

Es pot comprovar que per les mètriques més generals, MOTA i *idF1*, que els models de seguiment per detecció (SORT) és més encertat que la combinació de SORT amb filtres de correlació. Tot i això, és notable que la diferència no té per què ser significativa, ja que en el cas de deteccions per SDP la diferència per la MOTA i l'*idF1* és tan sols del 4,68% i del 0,18% respectivament. Aquesta proximitat en la bondat permetria seleccionar el model combinat en cas de requerir un rendiment elevant, per exemple, en sistemes de seguiment en temps real.

Detector	Model	Trajectòries	T. seguides	T. parcials	T. perdudes	#T. fragmen.
SDP	SORT	546	**147**	243	156	1405
SDP	combined	546	139	**270**	**137**	1655
YOLO	SORT	546	80	228	238	**1337**
YOLO	combined	546	71	240	235	1568

Taula 4.4: Nombre absolut de trajectòries en les escenes.

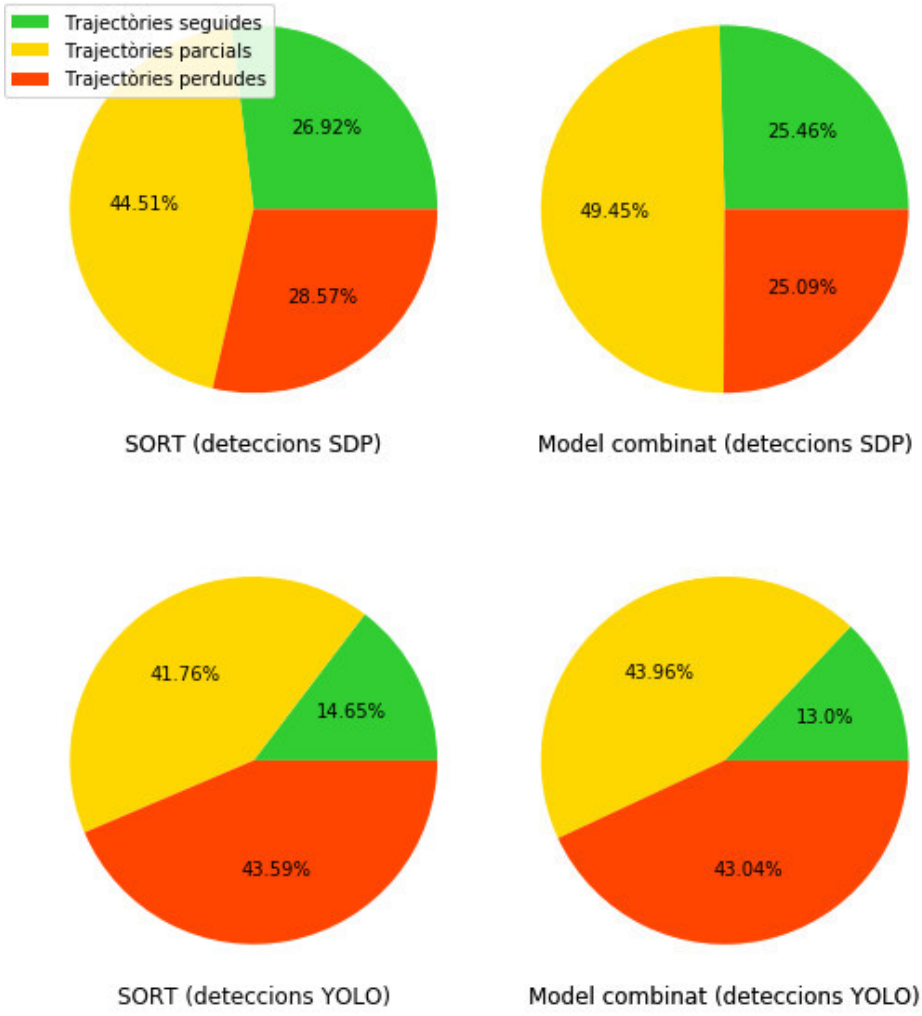

Figura 4.9: Distribució de trajectòries seguides, parcials i perdudes.

La taula 4.4 i la figura 4.9 mostren la distribució de les trajectòries trobades pels sistemes de seguiment respecte el *ground truth*. Aquestes mètriques, més tangibles que les ràtios anteriors, permeten una comprensió més intuïtiva del rendiment dels models pel que fa al seguiment dels objectes en moviment. Recordem que els conceptes de trajectòria seguida, parcial i perduda es corresponen a un encert de més del 80%, entre el 20%-80% i menys del 20% respectivament. S'observa com els resultats, novament, estan fortament lligats a la bondat de les deteccions. També resulta interessant observar com, tot i haver comprovat que les prediccions del model SORT són millors que les del model combinat entre SORT i els filtres de correlació, en ambdós casos redueixen el nombre de trajectòries perdudes. Aquest aspecte és rellevant a l'hora d'establir els llindars de qualitat per un sistema de seguiment: podria ser desitjable tenir un rendiment general més baix a costa d'evitar la perduda de trajectòria d'alguns casos.

4.3.3 Influència de les escenes

Una vegada comprovada la superioritat general de rendiment del model SORT, és oportú l'estudi desglossat per escenes. Com s'ha comprovat a la secció 4.2, les característiques de les escenes afecten directament a les deteccions, i per tant també ho faran al seguiment. Tot i això és interessant avaluar si existeixen diferències entre els models de seguiment que fan servir les mateixes deteccions com a dades d'entrada, o si les característiques visuals que dificulten la detecció també dificulten el seguiment.

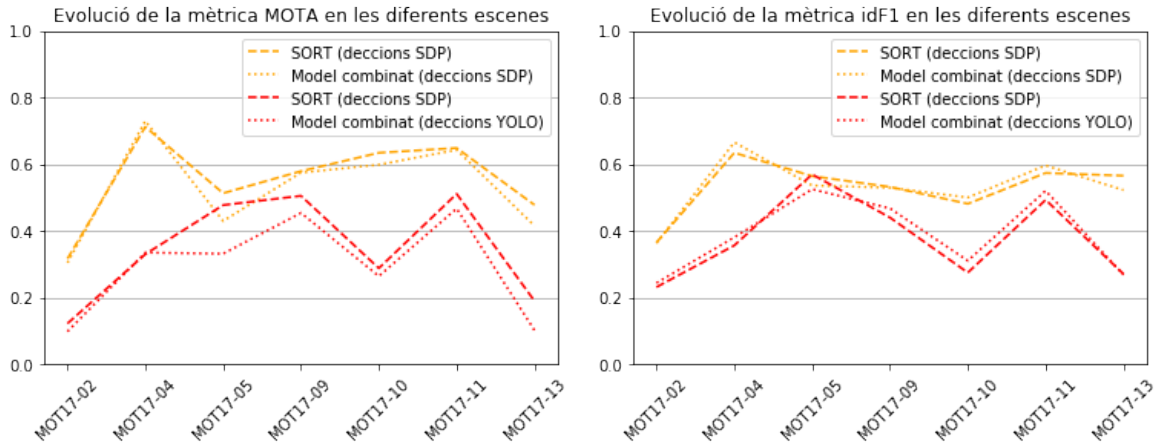

Figura 4.10: Comparativa entre escenes de les principals mètriques de seguiment.

A la figura 4.10 s'observen les principals mètriques dels models de seguiment en funció de l'escena tractada. Recordem que a la subsecció 4.2.1 s'ha mostrat com les escenes amb millor rendiment de detecció són les 05 i 11, mentre que les pitjors són les 02, 04 i 13. Si bé el rendiment relatiu es manté semblant entre les escenes, destaca especialment el cas de l'escena 04.

Pel que fa al rendiment de deteccions, l'escena 04 presenta dificultats, però té una bondat de seguiment superior a la mitjana. En aquesta escena, il·lustrada a la figura 4.11, es combinen dos factors que propicien aquest comportament: es produeixen aglomeracions de persones (el qual en dificulta la detecció de tots els individus), però s'enregistra des d'una perspectiva zenital. Aquesta orientació de la càmera evita moltes oclusions entre objectes, ja que els encreuaments entre vianants només oculten parcialment a la persona més llunyana.

4.3.4 Conclusions dels experiments de seguiment

En aquest segon d'experimentació bloc s'han consolidat certes idees ja identificades als experiments de detecció, com que la natura de l'escena afecta molt al rendiment del sistema.

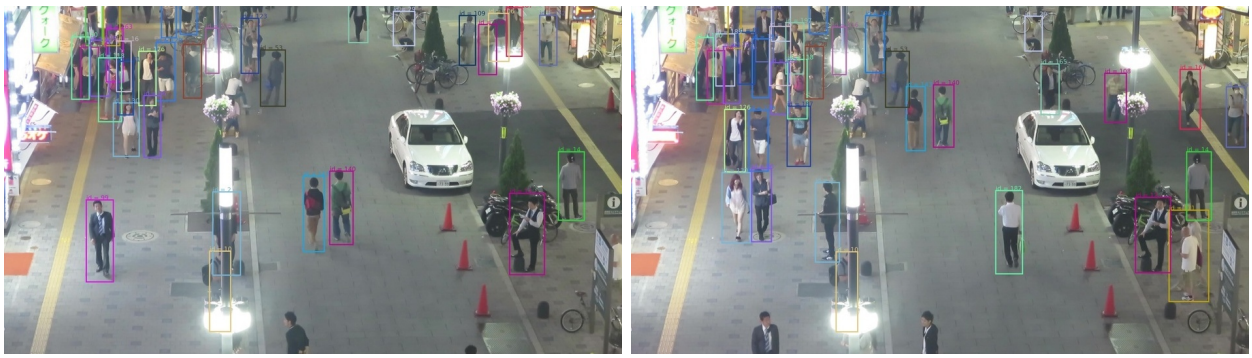

Figura 4.11: Seguiment SORT (amb deteccions SDP) a l'escena 04.

Pel que fa específicament als models de seguiment, d'entre les propostes presentades, el mètode SORT és el que millor efectivitat ofereix, però el veritablement rellevant ha sigut descobrir que la decisió d'escollir un sistema o un altre no només recau en les mètriques de bondat, sinó que es poden realitzar concessions de precisió en virtut d'altres aspectes com el cost computacional.

Finalment, cal remarcar la idea que als experiments presentats, els models de seguiments s'han vist fortament supeditats a la bondat de les deteccions, però que alguns aspectes es poden polir amb els *trackers* adequats.

Capítol 5

Conclusions

Una vegada conclosa la implementació i avaluació del projecte, cal fer la vista enrere i valorar diferents conclusions. Aquestes, pel caràcter acadèmic del projecte, van més enllà de les tècniques, i inclouen una perspectiva personal de l'alumne a tall de cloenda.

5.1 Resultat del projecte

Al llarg de les diferents etapes del projecte, des del capítol 2, amb l'estat de l'art, fins el capítol 4 amb l'anàlisi dels resultats s'ha vist clara una tendència: els problemes de detecció i seguiment són problemes oberts amb múltiples aproximacions per trobar-ne solució.

Concretament, s'ha identificat que les millors deteccions per les escenes MOTChallenge 17 s'han obtingut mitjançant un detector SDP i les millors mètriques de seguiment amb l'algorisme SORT, que es basa en el concepte de seguiment mitjançant detecció.

Si bé és cert que objectivament s'han trobat els millors models pel que fa a mètriques de bondat, també queda patent la idea que en funció del problema concret a resoldre, el científic de dades adquireix la responsabilitat d'identificar el millor model en funció de les característiques de les imatges amb les quals es treballarà.

Aquesta tasca de recerca no és gaire diferent de la resta de projectes d'aprenentatge automàtic, on és habitual realitzar una sèrie de fases de prototipat fins a trobar els models que millor s'adapten a la realitat del projecte concret.

5.2 Treball futur

Com bé s'ha pogut comprovar el repte de detecció i seguiment d'objectes està lluny de ser resolt completament. Per aquest motiu cal mantenir-se actualitzat amb les novetats que presenta la comunitat contínuament.

Com exemple de les aportacions més recents, en els darrers mesos s'han publicat novetats importants relacionades tant amb els reptes a resoldre com amb els models punters.

Pel que fa als reptes, el MOTChallenge ha publicat una nova edició amb escenes més complexes que l'edició treballada en aquest projecte [45]. Aquest nou repte presenta escenes amb molta densitat de vianants, que com s'ha comprovat és un dels factors que més afecta al rendiment dels sistemes de detecció. A més a més, també s'han ajustat les regions del *ground truth* per tal d'evitar objectes que no siguin persones.

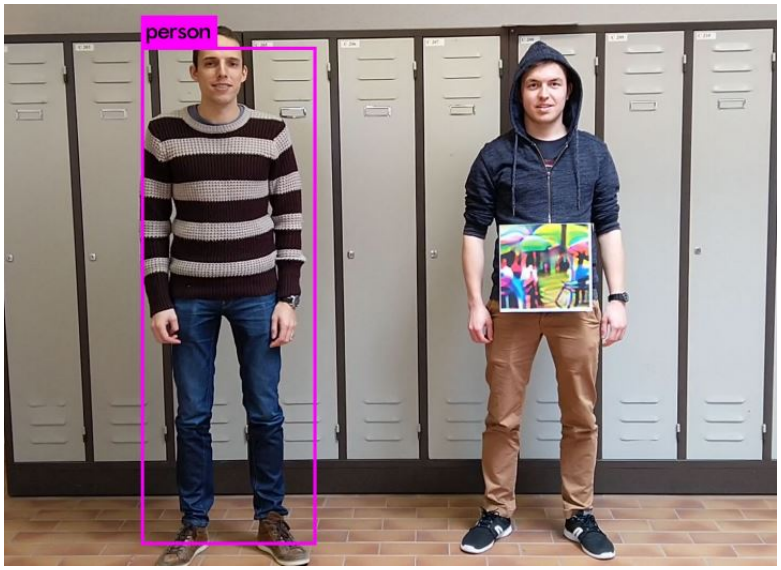

Figura 5.1: Mostra de la tècnica per evitar deteccions de YOLO.

Una altra novetat rellevant en relació a un dels models de detecció presentats, YOLO, és la creació d'una solució per evitar ser detectat com a vianant [86]. A la figura 5.1 s'observa com mitjançant un patró imprès subjectat per la persona, s'evita una detecció. Aquest experiment, més enllà de ser anecdòtic, presenta una sèrie de dilemes relatius als sistemes de vigilància ciutadana.

5.3 Lliçons apreses

A títol personal, i com a cloenda de la memòria del Treball Final, considero oportú exposar les diferents lliçons apreses durant l'execució del projecte.

Primerament, he trobat diferents barreres per a la comprensió del problema i les solucions actuals. El principal motiu d'aquests impediments és la vigència del problema i les constants innovacions que es donen. Per adquirir el coneixement necessari és imprescindible la lectura i comprensió d'articles científics, però també vull reconèixer la gran utilitat que suposen articles menys formals i de divulgació. Són una bona porta d'entrada a les àrees de coneixement. Com

a part de la referència bibliogràfica s'adjunten diferents blocs, tutorials o planes que m'han resultat especialment interessants [75, 55, 35].

A més de la barrera d'adquisició de coneixement, he hagut de superar la barrera tècnica. Aquesta, principalment s'ha degut a la dificultat per executar sistemes i models innovadors. Aquestes eines, diferents per cada gairebé experiment, requereixen un software [37].

Finalment, també cal indicar que un aprenentatge relatiu al procediment de treball basat en experiments, és que cal simplificar al màxim l'execució dels mateixos. És desitjable que els experiments siguin fàcilment repetibles, ja que per diferents avaluacions i comparatives és probable haver-los de re-executar.

Tot plegat, aquest Treball Final ha sigut una experiència enriquidora, no només pel coneixement tècnic adquirit, sinó també per la capacitat desenvolupada pel fet de realitzar un projecte relacionat completament amb tractament de dades, que té peculiaritats ben diferenciades de projectes clàssics de desenvolupament de software.

Acrònims

ANN *Artificial Neural Networks*. 8

AP *Average Precision*. 29

CamShift *Continuously Adaptive Mean Shift*. 11, 20, 21, 36

CNN *Convolutional Neural Networks*. 11, 17, 29

DPM *Deformable Part Models*. 15–17, 29, 30

EAO *Expected Average Overlap*. 11

ETL *Extract, Transform and Load*. 19, 25

F-RCNN *Faster RCNN*. 15

HoG *Histogram of Gradients*. 7, 8

IoU *Intersection over Union*. 9, 10, 19

MOT *Multiple Object Tracking*. 11, 12

MOTA *Multiple Object Tracking Accuracy*. 12, 37

R-CNN *Region-based Convolutional Neural Networks*. 8, 9, 18

SDP *Scale Dependent Pooling*. 15, 17, 30, 36, 37, 41

SORT *Simple Online and Realtime Tracking*. 23, 24, 36–41

SPP *Spatial Pyramid Pooling*. 8

SVM *Support Vector Machines*. 7, 11

VOT *Visual Object Tracking*. 11, 12

YOLO *You only Look once*. 9, 18, 23, 24, 29, 30, 36, 42

Bibliografia

[1] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online multiple instance learning. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 983–990. IEEE, 2009.

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 2008.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. February 2016.

[4] David S. Bolme, J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550, 2010.

[5] G. R. Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No.98EX201)*, pages 214–219, Oct 1998.

[6] Subhash Challa, Mark R. Morelande, Darko Mušicki, and Robin J. Evans. *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[7] cheind. cheind/py-motmetrics. https://github.com/cheind/py-motmetrics. Consultat: 2019-5-18.

[8] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. *CoRR*, abs/1504.02340, 2015.

[9] François Chollet et al. Applications - keras documentation. https://keras.io/applications/. Consultat: 2019-5-18.

[10] François Chollet et al. Keras. https://keras.io, 2015.

[11] Francois Chollet. *Deep Learning with Python*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2017.

[12] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, pages 142–149 vol.2, June 2000.

[13] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*, volume 1, pages 886–893, San Diego, United States, June 2005. IEEE Computer Society.

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[15] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. 2009.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[19] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes homepage. http://host.robots.ox.ac.uk/pascal/VOC/. Consultat: 2019-5-18.

[20] Daniel J. Fagnant and Kara Kockelman. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167 – 181, 2015.

[21] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. October 2017.

[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010.

[23] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, September 2010.

[24] Weitao Feng, Zhihao Hu, Wei Wu, Junjie Yan, and Wanli Ouyang. Multi-object tracking with multiple cues and switcher-aware classification. *CoRR*, abs/1901.06129, 2019.

[25] K. Gauen, R. Dailey, J. Laiman, Y. Zi, N. Asokan, Y. Lu, G. K. Thiruvathukal, M. Shyu, and S. Chen. Comparison of visual datasets for machine learning. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 346–355, Aug 2017.

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[27] Ross Girshick. Fast R-CNN. April 2015.

[28] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. November 2013.

[29] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[30] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. March 2017.

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.

[35] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. GluonCV: a deep learning toolkit for computer vision — gluoncv 0.5.0 documentation. https://gluon-cv.mxnet.io/. Consultat: 2019-5-18.

[36] John Hearty. *Advanced Machine Learning with Python*. Packt Publishing, 2016.

[37] Docker Inc. Enterprise application container platform — docker. https://www.docker.com/. Consultat: 2019-5-18.

[38] Davis King et al. dlib c++ library. image processing. http://dlib.net/imaging.html#correlation_tracker. Consultat: 2019-5-18.

[39] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pfugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, Gustavo Fernandez, and et al. The sixth visual object tracking vot2018 challenge results, 2018.

[40] Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin Zajc, Tomas Vojir, Gustav Häger, Alan Lukežič, Abdelrahman Eldesokey, and Gustavo Fernandez. The visual object tracking vot2017 challenge results, 2017.

[41] Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin Zajc, Tomas Vojir, Gustav Häger, Alan Lukežič, and Gustavo Fernandez. The visual object tracking vot2016 challenge results. Springer, Oct 2016.

[42] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016.

[43] Harold W Kuhn. The hungarian method for the assignment problem. *undefined*, 2010.

[44] Robert E Larson, Robert M Dressler, and Robert S Ratner. Application of the extended kalman filter to ballistic trajectory estimation. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1967.

[45] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Cvpr 2019 tracking challenge. https://motchallenge.net/data/CVPR_2019_Tracking_Challenge/, 2019. Consultat: 2019-6-02.

[46] Laura Leal-Taixé, Anton Milan, Ian D. Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *CoRR*, abs/1504.01942, 2015.

[47] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian D. Reid, and Stefan Roth. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *CoRR*, abs/1704.02781, 2017.

[48] Leal-Taixé, L. and Milan, A. and Reid, I. and Roth, S. and Schindler, K. MOT challenge. https://motchallenge.net/devkit. Consultat: 2019-5-18.

[49] A Li, M Lin, Y Wu, MH Yang, and S Yan. NUS-PRO: A New Visual Tracking Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):335–349, 2016.

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. May 2014.

[51] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[52] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[53] Liuliu. DPM: Deformable parts model. http://libccv.org/doc/doc-dpm/. Consultat: 2019-5-18.

[54] Wenhan Luo, Xiaowei Zhao, and Tae-Kyun Kim. Multiple object tracking: A review. *CoRR*, abs/1409.7618, 2014.

[55] Satya Mallick. MultiTracker : Multiple object tracking using OpenCV (C++/Python) — learn OpenCV. https://www.learnopencv.com/multitracker-multiple-object-tracking-using-opencv-c-python/, August 2018. Consultat: 2019-5-18.

[56] Santiago Manen, Michael Gygli, Dengxin Dai, and Luc Van Gool. Pathtrack: Fast trajectory annotation with path supervision. *CoRR*, abs/1703.02437, 2017.

[57] Paul Barham et al. Martín Abadi, Ashish Agarwal. Keras — TensorFlow core — TensorFlow. https://www.tensorflow.org/guide/keras. Consultat: 2019-5-18.

[58] Matterport. matterport/Mask_RCNN. https://github.com/matterport/Mask_RCNN. Consultat: 2019-5-18.

[59] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for Multi-Object tracking. March 2016.

[60] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[61] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[62] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[63] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[64] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas MOgelmose, Thomas B. Moeslund, and Sergio Escalera. Multi-modal rgb—depth—thermal human body segmentation. *Int. J. Comput. Vision*, 118(2):217–239, June 2016.

[65] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B. Moeslund, and Sergio Escalera. Multi-modal rgb–depth–thermal human body segmentation. *International Journal of Computer Vision*, 118(2):217–239, Jun 2016.

[66] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[67] Simon J. D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, New York, NY, USA, 1st edition, 2012.

[68] qqwweee. qqwweee/keras-yolo3. https://github.com/qqwweee/keras-yolo3. Consultat: 2019-5-18.

[69] rafaelpadilla. rafaelpadilla/Object-Detection-Metrics. https://github.com/rafaelpadilla/Object-Detection-Metrics. Consultat: 2019-5-18.

[70] rbgirshick. rbgirshick/voc-dpm. https://github.com/rbgirshick/voc-dpm. Consultat: 2019-5-18.

[71] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[72] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. April 2018.

[73] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[74] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. Beyond the kalman filter. *IEEE Aerospace and Electronic Systems Magazine*, 19(7):37–38, 2004.

[75] Adrian Rosebrock. PyImageSearch - be awesome at OpenCV, python, deep learning, and computer vision. https://www.pyimagesearch.com/. Consultat: 2019-5-18.

[76] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. September 2014.

[77] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *CoRR*, abs/1701.01909, 2017.

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for Large-Scale image recognition. September 2014.

[79] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, July 2014.

[80] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, pages 1–44, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[81] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The CLEAR 2006 evaluation. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 1–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[82] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. December 2015.

[83] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2010.

[84] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. August 2018.

[85] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-person tracking by multicut and deep matching. *CoRR*, abs/1608.05404, 2016.

[86] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. *CoRR*, abs/1904.08653, 2019.

[87] Luka Čehovin Zajc, Aleš Leonardis, and Matej Kristan. Visual object tracking performance measures revisited, Apr 2016.

[88] Y. Wu, J. Lim, and M. Yang. Online object tracking: A benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, June 2013.

[89] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, Sep. 2015.

[90] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[91] Hanxuan Yang, Ling Shao, Feng Zheng, Liang Wang, and Zhan Song. Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823 – 3831, 2011.

[92] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), December 2006.

[93] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. *CoRR*, abs/1711.06897, 2017.

[94] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. *CoRR*, abs/1811.04533, 2018.

[95] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 2019.

[96] ZidanMusk. ZidanMusk/experimenting-with-sort. https://github.com/ZidanMusk/experimenting-with-sort. Consultat: 2019-5-18.