
This is the **published version** of the bachelor thesis:

Souto Eslava, Gerard. *Automatització de la segmentació i adaptació de continguts audiovisuals a format vertical*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334379>

under the terms of the  license.

Automatització de la segmentació i adaptació de continguts audiovisuals a format vertical

Gerard Souto Eslava

29 de juny de 2026

Resum– L'increment del consum de contingut audiovisual en format vertical ha generat la necessitat d'adaptar vídeos originalment enregistrats en format horitzontal. Aquest treball presenta el desenvolupament d'un sistema automàtic de reenquadrament que permet convertir continguts 16:9 a format 9:16, mantenint la informació rellevant dins la zona visible.

La proposta combina tècniques de visió per computador per detectar persones i cares, fer-ne el seguiment temporal i generar automàticament l'enquadrament vertical, amb l'objectiu d'obtenir una composició coherent i estable. El sistema està orientat a reduir la intervenció manual en l'edició i facilitar l'adaptació de contingut a plataformes mòbils.

Els resultats mostren que el sistema genera reenquadraments automàtics útils, reduint significativament el temps d'edició respecte al procés manual i servint com a base per a ajustos posteriors.

Paraules clau– reenquadrament automàtic, processament de vídeo, detecció de persones i cares, seguiment temporal, detecció de canvis de pla.

Abstract– The increase in consumption of vertical video content has created the need to adapt videos originally recorded in horizontal format. This work presents the development of an automatic video reframing system that converts 16:9 content to 9:16 format while preserving relevant information within the visible area.

The proposed approach combines computer vision techniques to detect people and faces, perform temporal tracking, and automatically generate a vertical composition, aiming to achieve coherent and stable reframing. The system is designed to reduce manual editing effort and facilitate content adaptation for mobile platforms.

The results show that the system produces useful automatic reframing outputs, significantly reducing editing time compared to manual workflows while serving as a basis for further manual refinement.

Keywords– automatic reframing, video processing, person and face detection, temporal tracking, shot change detection.



1 INTRODUCCIÓ - CONTEXT DEL TREBALL

EN els darrers anys, el consum de contingut audiovisual a través de plataformes digitals com TikTok, Instagram Reels o YouTube Shorts ha experimentat un creixement molt significatiu. Aquestes plataformes utilitzen principalment el format vertical, fet que ha generat

la necessitat d'adaptar continguts produïts originalment en format horitzontal 16:9 a relacions d'aspecte verticals 9:16.

Actualment, aquest procés encara depèn en gran mesura de la intervenció manual, especialment en la selecció de regions rellevants i l'ajust dinàmic de l'enquadrament al llarg del vídeo. Això comporta un cost elevat en temps i recursos, especialment quan s'ha de processar una gran quantitat de contingut audiovisual.

Els avenços en visió per computador i processament de vídeo permeten automatitzar part d'aquest procés mitjançant tècniques de detecció i seguiment de subjectes, detecció de canvis de pla i anàlisi de regions d'interès. La combinació d'aquestes tècniques permet desenvolupar sistemes capaços d'identificar automàticament el contingut re-

- E-mail de contacte: gerardsouto33@gmail.com
- Menció: Computació
- Treball tutoritzat per: Lluís Gómez (Ciències de la Computació)
- Curs 2025/26

llevant i generar reenquadraments coherents adaptats al format vertical.

Aquest treball s'ha desenvolupat en el marc de les pràctiques realitzades a 3Cat, dins d'un entorn professional del sector audiovisual. El projecte respon a una necessitat real de l'empresa relacionada amb l'adaptació automàtica de continguts horitzontals a formats verticals per a la seva distribució en plataformes digitals.

En aquest context, es presenta un sistema automàtic de reenquadrament per a la conversió de vídeos horitzontals a format vertical, amb l'objectiu de reduir la intervenció manual i millorar l'eficiència del procés d'adaptació audiovisual.

1.1 Objectius

L'objectiu principal d'aquest treball és desenvolupar un sistema automàtic capaç d'adaptar vídeos en format horitzontal 16:9 a format vertical 9:16, mantenint dins de la regió visible els elements rellevants de cada escena.

Per assolir aquest objectiu general es defineixen els objectius específics següents:

- Analitzar les tècniques existents de reenquadrament automàtic, detecció de canvis de pla i identificació de regions d'interès en vídeo.
- Dissenyar l'arquitectura general del sistema definint els mòduls necessaris per al procés de reenquadrament automàtic.
- Desenvolupar un mètode de detecció automàtica de canvis de pla per segmentar el vídeo en escenes coherents.
- Desenvolupar un sistema d'anàlisi visual que permeti identificar els principals elements d'interès presents en cada escena.
- Desenvolupar un mètode capaç de determinar automàticament l'enquadrament vertical més adequat a partir de la informació visual obtinguda.
- Avaluar qualitativament i quantitativament els resultats obtinguts i comparar-los amb el procés d'edició manual.

1.2 Estructura del document

La resta del document s'organitza de la manera següent. En primer lloc, la secció d'estat de l'art revisa els principals treballs relacionats amb el reenquadrament automàtic de vídeo, la detecció de canvis de pla i el seguiment d'objectes.

A continuació es presenta la metodologia utilitzada i l'arquitectura del sistema desenvolupat.

Posteriorment s'exposen els resultats obtinguts i la seva anàlisi, incloent una comparació amb l'enfocament d'edició manual.

Finalment es recullen les conclusions del treball, així com possibles línies de millora i treball futur.

2 ESTAT DE L'ART

El problema del reenquadrament automàtic de vídeo ha rebut una atenció creixent en els darrers anys, especialment en el context de l'adaptació de continguts audiovisuals a dispositius mòbils i plataformes de consum vertical. Aquest problema es pot descompondre en diversos subproblemes principals, com ara la segmentació temporal del vídeo, la detecció de regions d'interès i el seguiment coherent d'objectes al llarg del temps.

2.1 Segmentació temporal i detecció de canvis de pla

La segmentació de vídeo en escenes és un pas fonamental per a qualsevol sistema de processament temporal. Les tècniques clàssiques es basen en la comparació entre fotogrames consecutius mitjançant diferències d'histogrames o mesures de distància en espais de color [2], permetent detectar canvis abruptes de contingut visual i separar el vídeo en plans coherents [4].

En aquest context, l'ús de diferents espais de color, com RGB o HSV, ha demostrat ser útil per millorar la robustesa de la detecció davant variacions d'il·luminació [3]. No obstant això, aquests mètodes poden ser sensibles a canvis graduals o a moviments de càmera, generant falsos positius en determinades situacions. Eines com PySceneDetect implementen aquestes estratègies i s'utilitzen habitualment en aplicacions pràctiques de segmentació de vídeo [1].

2.2 Reenquadrament automàtic de vídeo

El reenquadrament automàtic té com a objectiu seleccionar, per a cada instant del vídeo, la regió més rellevant per ser mostrada en un format vertical. Tradicionalment, aquest problema s'ha abordat mitjançant tècniques de reenquadrament basades en el contingut, on es calcula una regió d'interès a partir de característiques visuals com el moviment, el contrast o la presència d'objectes, amb l'objectiu de maximitzar la informació visual dins del nou enquadrament [8].

Els enfocaments més recents utilitzen models d'aprenentatge profund per estimar la rellevància espacial i temporal de les diferents regions de la imatge, incorporant informació semàntica que permet identificar de manera més precisa les regions visualment més rellevants, especialment en escenes complexes [5, 6]. No obstant això, aquests mètodes poden presentar limitacions en la coherència temporal del reenquadrament, així com dependència de dades d'entrenament específiques i dificultats en la generalització a contingut molt variat.

En aquest àmbit, diversos treballs han proposat sistemes complets de reenquadrament automàtic per a formats verticals, obtenint bons resultats en contingut dinàmic, tot i que el rendiment depèn de la complexitat de l'escena, la presència de múltiples subjectes o moviments de càmera ràpids [7, 8].

2.3 Detecció i seguiment d'objectes

Un dels elements clau en els sistemes de reenquadrament automàtic és la capacitat d'identificar els subjectes visualment més rellevants de cada escena. En molts casos, les

persones i les seves cares constitueixen el principal focus d'atenció, fet que converteix la seva detecció en una tasca fonamental dins del procés d'anàlisi del vídeo.

Les tècniques actuals de detecció d'objectes es basen en models d'aprenentatge profund capaços de localitzar diferents categories amb alta precisió i en temps gairebé real. Entre aquests destaquen variants de YOLO per a la detecció de persones i altres objectes [17], així com detectors facials com RetinaFace, dissenyats per ser robustos davant variacions d'il·luminació, posició i oclusions parcials [11].

Per mantenir la coherència temporal, s'utilitzen mecanismes de seguiment que associen deteccions entre fotogrames consecutius i permeten construir trajectòries consistents, com SORT o DeepSORT [12, 13]. En escenes complexes, aquests mètodes es poden complementar amb tècniques de reidentificació basades en característiques visuals o facials per millorar l'estabilitat del seguiment [15, 16].

2.4 Síntesi de l'estat de l'art

En conjunt, la literatura analitzada indica que el reenquadrament automàtic de vídeo depèn de la combinació de segmentació temporal, detecció d'objectes i seguiment per garantir la coherència de les regions d'interès.

Tot i els avenços existents, encara persisteixen dificultats en escenes complexes, amb canvis bruscos de càmera o múltiples subjectes en moviment.

En aquest context, aquest treball proposa una aproximació basada en la integració d'aquestes tècniques per al reenquadrament automàtic de vídeo en format vertical.

3 METODOLOGIA I PROCÉS DE DESENVOLUPAMENT

Aquest capítol descriu la metodologia seguida durant el desenvolupament del treball, així com les diferents fases d'implementació del sistema. També es presenten les decisions tècniques adoptades, l'evolució dels principals mòduls i les proves realitzades per validar-ne el funcionament.

3.1 Metodologia de treball

El desenvolupament del projecte s'ha dut a terme seguint un enfocament incremental i experimental, que ha permès abordar el problema de manera progressiva i validar els diferents components abans de la seva integració en el sistema final.

En una primera fase es va realitzar una revisió bibliogràfica de les tècniques relacionades amb la detecció de canvis de pla, la detecció i seguiment de cares i persones, i els sistemes de reenquadrament automàtic de vídeo. Aquesta anàlisi va permetre identificar les solucions existents i establir les bases tècniques del projecte.

A partir d'aquesta base es va definir l'arquitectura general del sistema i es van identificar els principals mòduls funcionals. El desenvolupament es va organitzar mitjançant un procés iteratiu basat en la metodologia àgil Scrum.

La gestió de les tasques es va realitzar amb Jira, utilitzant-lo com a eina de planificació i seguiment de les

diferents iteracions del projecte. Paral·lelament, es va utilitzar Git com a sistema de control de versions per gestionar l'evolució del codi i facilitar la integració dels canvis.

Cada iteració seguia un cicle format per la implementació dels algorismes seleccionats, la validació dels resultats obtinguts i l'anàlisi dels problemes detectats, permetent una evolució progressiva del sistema.

3.2 Planificació del projecte

La planificació del TFG es va estructurar en diverses fases que cobreixen el desenvolupament complet del sistema, des de la definició inicial dels requisits fins a la redacció de la memòria i la preparació de la defensa final. Aquesta proposta va servir com a guia per organitzar les tasques de recerca, desenvolupament, integració i validació del sistema.

La Figura 1 mostra el diagrama de Gantt utilitzat per planificar temporalment les diferents activitats del projecte. Les fases principals inclouen la revisió de l'estat de l'art, el disseny de l'arquitectura del sistema, la implementació dels diferents mòduls, la integració i validació del sistema, i la redacció de la memòria final.

En termes generals, la planificació inicial s'ha complert amb ajustos menors derivats del procés iteratiu de desenvolupament, sense afectar significativament el calendari global del projecte.

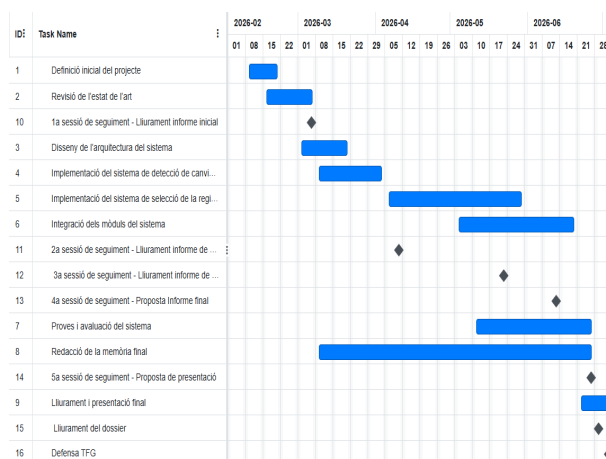


Fig. 1: Diagrama de Gantt amb la planificació temporal de les diferents fases del projecte.

3.3 Arquitectura general del sistema

La Figura 2 mostra l'arquitectura general del sistema desenvolupat. El sistema segueix un disseny modular basat en un flux de processament de vídeo, en què es transformen seqüencialment els fotogrames d'entrada fins a obtenir una versió reenquadrada en format vertical.

L'arquitectura es divideix en tres blocs principals. En primer lloc, una etapa d'entrada on l'usuari configura els paràmetres i inicia el procés de processament a través de la interfície del sistema. A continuació, el nucli del sistema (mòdul de processament) s'encarrega de l'anàlisi del vídeo, que inclou la segmentació en escenes, la detecció d'elements d'interès i el càlcul de les ROI (Region Of Interest). Finalment, un mòdul de sortida genera el vídeo reenquadrat en format vertical juntament amb la informació

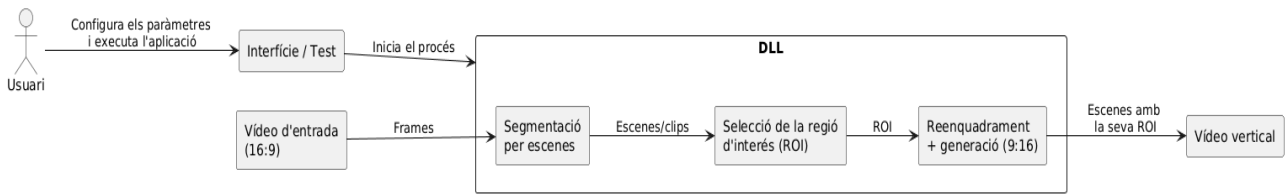


Fig. 2: Arquitectura general del sistema de reenquadrament automàtic de vídeo.

associada a les ROIs calculades.

Aquest disseny modular permet desacoblar les diferents funcionalitats del sistema, facilitant la seva extensió i manteniment, així com la possible substitució de components específics.

3.4 Disseny i implementació del sistema

A continuació es descriuen els principals mòduls implementats durant el desenvolupament del projecte, així com les diferents iteracions realitzades fins a arribar a la versió final del sistema.

3.4.1 Entorn de desenvolupament

La implementació s'ha realitzat en llenguatge C++, una decisió condicionada per l'entorn de desenvolupament de l'empresa i pels requisits de rendiment en el processament de vídeo en temps real. Per al desenvolupament de les funcionalitats de visió per computador s'ha utilitzat la biblioteca OpenCV [18], que proporciona les eines necessàries per al tractament de vídeo, la detecció d'objectes i el processament d'imatges de manera eficient.

L'arquitectura del sistema s'ha implementat mitjançant una biblioteca dinàmica (DLL), seguint l'estàndard del sistema existent a l'empresa, amb l'objectiu de facilitar la integració del mòdul desenvolupat dins d'una plataforma més àmplia i modular. Per a la validació del sistema s'ha desenvolupat una aplicació de test basada en Microsoft Foundation Class Library (MFC), que permet executar el sistema amb vídeos reals, visualitzar els resultats i ajustar els paràmetres de funcionament.

El desenvolupament s'ha dut a terme en l'entorn Microsoft Visual Studio sobre sistema operatiu Windows.

3.4.2 Detecció de canvis de pla

La detecció de canvis de pla constitueix la primera etapa del sistema, amb l'objectiu d'identificar transicions entre plans i dividir el vídeo en fragments coherents.

Durant el desenvolupament es van analitzar diferents aproximacions. Inicialment es va implementar un mètode basat en la diferència absoluta entre fotogrames consecutius, sensible a moviments de càmera i dels subjectes.

La versió final es basa en la comparació d'histogrames de color en espai HSV [3]. Cada fotograma es divideix en regions espacials i, per a cadascuna, es calculen histogrames dels canals Hue i Saturation, normalitzats per garantir comparació consistent [2]. La similitud entre histogrames es mesura mitjançant la distància de Bhattacharyya [4]:

$$D(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{H_1 H_2}} \sum_{i=1}^N \sqrt{H_1(i) H_2(i)}} \quad (1)$$

Els valors obtinguts s'agreguen globalment per estimar la variació entre fotogrames, i es detecta un canvi de pla quan se supera un llindar predefinit. Els histogrames del fotograma actual es conserven per a la comparació amb el següent.

3.4.3 Detecció de cares, seguiment i càlcul de la ROI

La detecció de cares constitueix el mecanisme principal utilitzat pel sistema per identificar els subjectes d'interès i determinar el reenquadrament del vídeo. Per a aquesta tasca es van incorporar dos detectors facials alternatius: el detector basat en Deep Neural Network (DNN) d'OpenCV [9] i el model YuNet [10]. Tot i que YuNet ofereix una major precisió i robustesa en la detecció facial, especialment en escenes més complexes, el detector DNN presenta un menor cost computacional. Per aquest motiu, el sistema permet seleccionar el detector més adequat segons les característiques del vídeo, utilitzant habitualment YuNet com a opció preferent i reservant el detector DNN per a continguts on les cares són grans i fàcilment identificables.

Durant les primeres iteracions, el càlcul de la regió d'interès (ROI) es basava únicament en la cara de major mida detectada en cada fotograma. Tot i obtenir resultats acceptables en escenes senzilles, aquesta aproximació generava inestabilitats quan apareixien múltiples persones a l'escena o es produïen pèrdues temporals de detecció.

Per millorar la coherència temporal es va implementar un mecanisme de seguiment persistent de cares. El sistema assigna una identitat a cada persona detectada combinant criteris de proximitat espacial amb la comparació de descriptors facials (embeddings), representacions numèriques de cada rostre, generats mitjançant un model de reconeixement facial basat en ArcFace [15]. Aquesta estratègia permet mantenir la identificació dels subjectes entre fotogrames consecutius i afavoreix la seva recuperació després d'oclusions o desaparicions temporals.

La Figura 3 mostra un exemple del sistema de detecció i seguiment facial durant l'execució.

A partir de les identitats detectades es calcula la ROI associada a cada fotograma. Addicionalment, es va incorporar un mecanisme de *lookahead* basat en l'anàlisi de fotogrames futurs, que permet validar la persistència dels subjectes abans de fixar el reenquadrament. Aquest enfocament redueix l'impacte de deteccions puntuals incorrectes i contribueix a obtenir trajectòries més estables.

També es van implementar tècniques de suavització temporal i limitació del moviment de la ROI per evitar



Fig. 3: Exemple de detecció facial i seguiment de múltiples subjectes mitjançant identitats persistents.

desplaçaments bruscos entre fotogrames consecutius, així com un sistema de zoom configurable segons la dimensió aparent de les cares detectades.

Finalment, el sistema incorpora suport per a escenes amb múltiples persones. En aquests casos, s'analitza la persistència temporal de les diferents identitats detectades i es seleccionen aquelles més consistents al llarg de l'escena, permetent calcular reenquadraments més robustos i adequats al contingut visual.

Algorithm 1 FaceDetectorDNN: detecció, tracking i càlcul de ROIs

```

1: Inicialitzar model facial (DNN o YuNet) i embeddings
2: for cada frame do
3:   Detectar cares
4:   Extreure embeddings facials
5:   for cada cara detectada do
6:     Intentar associar ID:
7:       - per proximitat espacial
8:       - per similitud d'embeddings
9:   if no coincidència then
10:     Crear nova identitat
11:   end if
12: end for
13: end for
14: Analitzar seqüència completa de frames (lookahead)
15: for cada frame do
16:   Determinar cara principal o identitat dominant
17:   Comprovar estabilitat en frames futurs
18:   if identitat estable then
19:     Calcular ROI centrada en la cara
20:   else if no hi ha estabilitat inicial then
21:     Buscar identitat dominant en frames futurs
22:     Utilitzar-la com a referència
23:   else
24:     Reutilitzar ROI anterior
25:   end if
26:   if mode transició limitat then
27:     Aplicar restriccions de moviment i zoom
28:   end if
29:   Guardar ROI del frame
30: end for
31: Retornar seqüència de ROIs temporals

```

3.4.4 Detecció de persones

Per complementar la detecció facial, es va incorporar un mòdul de detecció de persones basat en un model de xarxa neuronal en format ONNX (YOLOv5n) [17]. Aquest component és opcional i s'utilitza en escenes on es prioritza la preservació del fotograma complet, com en continguts amb presentadors o situacions on el context global és rellevant.

El model s'executa sobre cada fotograma i genera deteccions de la classe "persona", que es filtren mitjançant un llindar de confiança, Non-Maximum Suppression i criteris de mida per descartar deteccions no representatives. La detecció amb major àrea es considera la més rellevant.

A partir de les deteccions obtingudes, el sistema avalua la persistència temporal de la presència humana. Si es detecta una presència estable i la configuració ho indica, es manté el fotograma complet com a regió d'interès. En cas contrari, el sistema delega el càlcul de la ROI al mòdul de detecció facial.

En aquest sentit, aquest mòdul actua com un criteri de selecció superior que permet alternar entre reenquadrament i preservació del frame complet segons el contingut de l'escena.

Algorithm 2 PersonDetector: detecció de persones

```

1: Carregar model ONNX
2: for cada frame do
3:   Executar inferència
4:   Filtrar deteccions de persona
5:   Aplicar NMS i confiança
6:   if no hi ha detecció vàlida then
7:     Retornar "no detectat"
8:   else
9:     Seleccionar detecció principal
10:    Validar mida respecte el frame
11:   end if
12: end for
13: Retornar persona detectada

```

3.4.5 Integració dels mòduls

La integració dels diferents components del sistema es realitza mitjançant la classe `VideoAnalyzer`, que actua com a controlador principal del flux d'anàlisi de vídeo.

Aquest mòdul centralitza i coordina totes les etapes del sistema, incloent la segmentació del vídeo en escenes, la detecció de cares i persones, el càlcul de les regions d'interès (ROI) i la generació dels resultats finals. D'aquesta manera, es garanteix una execució seqüencial i coherent de tots els components desenvolupats de forma independent.

En funció del contingut de cada escena i de les deteccions disponibles, el sistema decideix dinàmicament l'estratègia de reenquadrament a aplicar, prioritzant la detecció de persones quan és fiable i utilitzant la informació facial com a base principal en la resta de casos.

Finalment, el mòdul incorpora funcionalitats d'exportació dels resultats en format JSON, facilitant la traçabilitat dels experiments i la integració amb eines externes d'anàlisi.

Els Algorismes 1, 2 i 3 resumeixen el funcionament dels principals components del sistema.

Algorithm 3 VideoAnalyzer: anàlisi del vídeo i càlcul de ROIs

```

1: Inicialitzar paràmetres i obrir vídeo
2: Obtenir FPS, dimensions i rang de frames
3: Inicialitzar detectors (escenes, cares i persones si escau)
4: Detectar intervals d'escena
5: for cada escena do
6:   Carregar frames de l'escena
7:   if detecció de persones activada i persona dominant detectada then
8:     Assignar ROI = frame complet
9:   else
10:    if múltiples cares activades then
11:      Calcular múltiples ROIs amb lookahead
12:    else
13:      Calcular ROI única amb lookahead
14:    end if
15:  end if
16:  Guardar ROIs de l'escena
17: end for
18: Exportar resultats a JSON
19: Alliberar recursos

```

3.5 Validació i proves

El funcionament del sistema s'ha validat mitjançant diferents entorns de prova específics per a cada component, basats en l'anàlisi visual dels resultats i l'ajust iteratiu dels paràmetres.

En primer lloc, es va validar el mòdul de detecció de canvis de pla, analitzant la precisió de la segmentació en diferents tipus de contingut.

Posteriorment, es va avaluar el càlcul de les regions d'interès i la detecció de cares, amb atenció a l'estabilitat temporal del reenquadrament.

Finalment, es va realitzar un test integral de la classe VideoAnalyzer per verificar la correcta integració de tots els mòduls.

L'entorn de proves utilitzat per a la validació del sistema es mostra a la Figura 4, on es poden observar els diferents panells de control i visualització utilitzats durant l'execució del sistema.

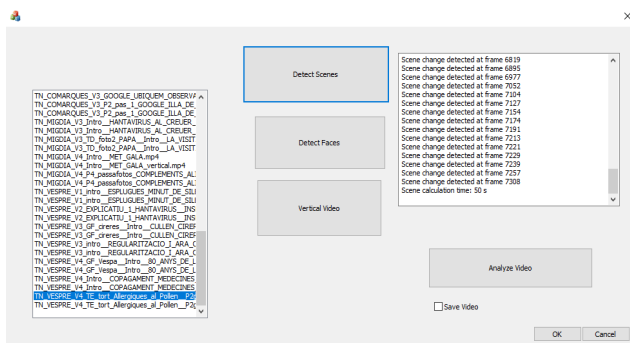


Fig. 4: Interfície de l'entorn de proves utilitzat per a la validació del sistema.

4 RESULTATS I AVALUACIÓ

4.1 Resultats quantitatius

4.1.1 Detecció de canvis de pla

S'han realitzat experiments sobre diferents vídeos reals amb l'objectiu d'avaluar el rendiment de l'algorisme de detecció de canvis de pla basat en histogrames HSV per blocs, tal com es mostra a la Figura 5. En aquest context, els valors de referència (*Reals*) corresponen al nombre de canvis de pla anotats manualment per a cada vídeo, els quals s'utilitzen com a *ground truth*.

Els resultats obtinguts es resumeixen a la Taula 1.

Vídeo	Reals	Detectats	Observacions
1	7	7	Plató estable. Resultat correcte.
2	7	5	Subdetecció en transicions suaus.
3	17	17	Correcte amb petites desviacions en zooms.
4	32	26	Subdetecció en plans similars.
5	33	32	Molt bon rendiment general.
6	19	19	Estructura estable. Correcte.
7	31	34	Sobresegmentació per moviment intens.
8	39	55	Sobresegmentació per llums i concerts.

TAULA 1: RESULTATS DE DETECCIÓ DE CANVIS DE PLA AMB L'ALGORISME HSV

A partir dels resultats obtinguts, es calcula una desviació absoluta mitjana de 3.25 canvis de pla per vídeo respecte als valors de referència, la qual reflecteix un comportament globalment estable del sistema.

En termes generals, l'algorisme presenta un bon rendiment en vídeos amb canvis de pla ben definits i escenes estables (Vídeos 1, 3 i 6), on la coincidència amb el ground truth és pràcticament exacta.

En canvi, en escenes amb transicions suaus o amb alta similitud visual entre plans (Vídeos 2 i 4), es produeix una tendència a la subdetecció, mentre que en continguts amb moviment intens o variacions d'il·luminació (Vídeos 7 i 8) s'observa sobresegmentació.

En comparació amb l'algorisme inicial basat en diferències absolutes entre fotogrames, el mètode basat en histogrames HSV millora la robustesa davant moviments globals de càmera i subjectes, reduint falsos positius en escenes estàtiques. No obstant això, manté limitacions en transicions suaus o amb canvis progressius de contingut, on la separació entre plans no és clarament definida.

4.1.2 Rendiment computacional i comparació amb edició manual

S'ha avaluat el temps de processament del sistema complet i s'ha comparat amb la duració dels vídeos originals, així com amb el temps estimat de conversió manual per part d'operadors d'edició professional.

Els resultats es presenten separatament segons la configuració utilitzada per al mòdul de detecció facial. En concret, es distingeixen dos escenaris: un primer grup de vídeos



Fig. 5: Exemple de detecció de canvis de pla amb visualització del número d'escena en temps real.

processats amb el model de detecció facial d'OpenCV, i un segon grup utilitzant el model YuNet.

En ambdós casos, el temps de processament depèn principalment del model de detecció facial emprat, que és el component més costós del pipeline. Addicionalment, poden influir altres paràmetres del sistema, com la detecció de persones o l'ús de mecanismes de lookahead, mentre que el nombre d'escenes del vídeo no presenta un impacte significatiu directe en el temps total.

Tots els experiments s'han executat en entorn CPU, sense acceleració per GPU, fet que incrementa els temps de processament respecte a una possible implementació optimitzada amb suport per GPU ja disponible a OpenCV, sempre que es disposi del maquinari adequat.

Els resultats corresponents al model d'OpenCV es mostren a la Figura 6, mentre que els resultats per al model YuNet es presenten a la Figura 7.

En tots els casos, el sistema requereix un temps de processament superior a la duració dels vídeos originals, fet esperable tenint en compte la complexitat del processament.

En l'entorn d'edició de 3Cat, la conversió manual d'un vídeo horitzontal a format vertical requereix aproximadament entre 25 i 35 minuts per cada 3 minuts de contingut, depenent de la complexitat de l'escena.

Tot i això, el principal avantatge del sistema és la reducció significativa del temps global del procés, ja que proporciona una base automàtica de reenquadrament que pot ser posteriorment ajustada per l'editor si és necessari.

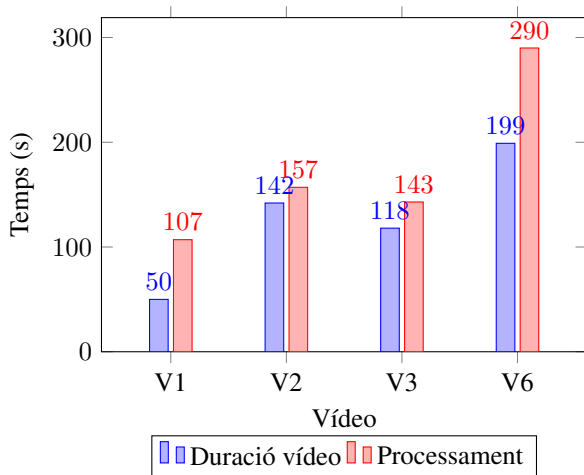


Fig. 6: Comparació de temps per a vídeos amb configuració del model d'OpenCV

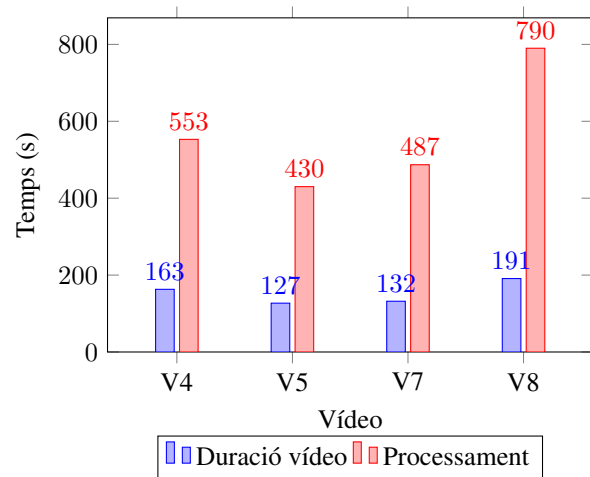


Fig. 7: Comparació de temps per a vídeos amb configuració del model YuNet

4.2 Resultats qualitatius

4.2.1 Verticalització de vídeos

Per complementar l'anàlisi quantitativa, s'ha realitzat una avaluació qualitativa del sistema a partir de 12 exemples representatius de vídeos reals. Per a cada cas es compara un fotograma original en format horitzontal amb el resultat obtingut en format vertical.

Els resultats qualitatius obtinguts es resumeixen a la Figura 8.

Els resultats mostren que el sistema genera reenquadraments coherents en la majoria d'escenaris amb presència de persones, especialment en continguts de plató, entrevistes i escenes informatives. En aquests casos, el sistema aconsegueix mantenir els subjectes principals dins del marc vertical de manera estable.

En escenes amb múltiples elements visuals o fons rellevants, el sistema prioritza la detecció de cares, fet que pot provocar la pèrdua d'informació contextual (pantalles, gràfics o elements de fons).

En escenes més dinàmiques o poc clares semànticament, s'observen limitacions en la selecció del subjecte principal, especialment quan aquest no coincideix amb el centre visual de l'escena o quan hi ha múltiples persones, ja que el sistema no incorpora informació sobre qui està parlant i es basa exclusivament en criteris visuals de detecció facial.

Finalment, en continguts esportius com futbol i tennis, el sistema presenta dificultats, ja que el reenquadrament basat en detecció facial no sempre captura l'acció principal.

Aquest conjunt d'exemples permet identificar tant els punts forts del sistema com les seves limitacions, especialment derivades de la dependència de la detecció facial com a criteri principal de reenquadrament.

4.3 Discussió global

Els resultats obtinguts mostren que el sistema és capaç de generar segmentacions temporals i reenquadraments coherents en una part important dels escenaris analitzats, especialment en continguts amb estructura estable i presència de subjectes clarament identificables, com ara platós i entrevistes.

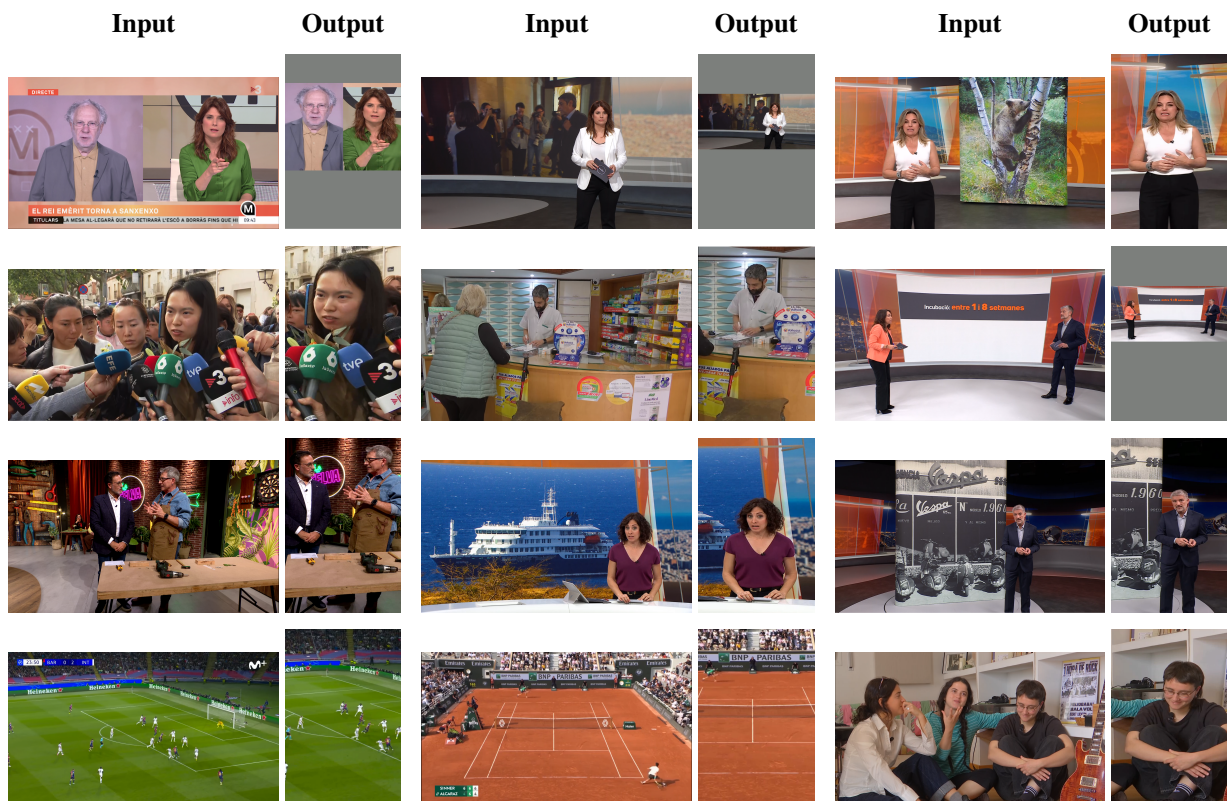


Fig. 8: Resultats qualitatius del sistema de reenquadrament

En la detecció de canvis de pla, l'ús d'histogrames HSV millora la robustesa respecte a mètodes basats en diferències de píxels, reduint la sensibilitat al moviment global de l'escena. Tot i això, persisteixen errors en transicions suaus o escenes visualment similars.

Pel que fa al reenquadrament, la combinació de detecció facial i seguiment temporal permet obtenir resultats estables en la majoria de casos, tot i que en escenes amb múltiples focus d'atenció o amb ambigüitat semàntica la selecció del subjecte principal pot no ser sempre la més adequada.

Finalment, el rendiment computacional depèn principalment del model de detecció facial utilitzat, mentre que la resta de components tenen un impacte secundari. En conjunt, el sistema es presenta com una eina viable de suport per a la postproducció automatitzada de contingut audiovisual.

5 CONCLUSIONS

Aquest treball ha permès dissenyar i implementar un sistema automàtic orientat a la conversió de contingut audiovisual horitzontal a format vertical mitjançant tècniques de visió per computador. La principal aportació del TFG ha estat la definició i desenvolupament d'una arquitectura modular capaç d'integrar segmentació en plans, detecció de cares i persones, càlcul dinàmic de regions d'interès i generació automàtica de reenquadraments adaptats al format vertical.

Pel que fa als objectius plantejats, s'ha assolit l'objectiu general de construir un sistema funcional capaç d'analitzar un vídeo i generar reenquadraments verticals visualment coherents en diversos escenaris. També s'han assolit els objectius específics relacionats amb la implementació del mòdul de detecció de canvis de pla, el desenvolupament

del sistema de selecció de regions d'interès i la integració dels diferents components en una única solució.

Els experiments realitzats han permès validar el comportament del sistema sobre diferents tipus de contingut audiovisual. Els resultats mostren que l'algorisme de detecció de canvis de pla basat en histogrames HSV és capaç d'obtenir resultats propers als valors reals en la majoria dels casos analitzats, mentre que el sistema de reenquadrament genera resultats visualment satisfactoris en escenaris amb subjectes clarament identificables. Així mateix, el procés complet s'ha automatitzat amb èxit, permetent obtenir una primera versió vertical del vídeo en un temps notablement inferior al requerit en un procés d'edició manual tradicional.

Entre les aportacions concretes del treball destaquen la comparació i selecció d'un mètode robust de detecció de canvis de pla, la incorporació de mecanismes de persistència temporal i *lookahead* per estabilitzar la regió d'interès, el suport a diferents detectors facials, la detecció complementària de persones de cos complet i la construcció de la classe integradora *VideoAnalyzer*, encarregada de coordinar el flux complet d'anàlisi i processament.

Tanmateix, el treball també ha posat de manifest algunes limitacions. La selecció de la regió d'interès continua depenent principalment de criteris visuals, fet que dificulta identificar el subjecte narrativament més rellevant. Aquesta limitació es fa especialment evident en escenes amb múltiples persones, on el sistema no disposa d'informació sobre qui està intervenint o parlant en cada moment. En conseqüència, tot i que el sistema constitueix una eina de suport útil, encara no pot substituir completament el criteri editorial humà en tots els escenaris possibles.

Pel que fa a les línies de treball futures, seria interessant complementar l'anàlisi visual amb informació semàntica i d'àudio, incorporant mecanismes de detecció del parlant

actiu o models especialitzats segons el tipus de contingut. També seria rellevant aprofundir en l'optimització del rendiment computacional i en la definició de mètriques objectives que permetin avaluar de manera més sistemàtica la qualitat del reenquadrament generat.

En síntesi, el TFG demostra que és possible automatitzar una part significativa del procés d'adaptació de vídeo horitzontal a format vertical utilitzant recursos i eines accessibles dins l'àmbit de l'enginyeria informàtica. El sistema desenvolupat ofereix resultats satisfactoris en una àmplia varietat d'escenaris, i estableix una base sòlida per continuar avançant cap a sistemes de reenquadrament automàtic més intel·ligents i adaptats a entorns reals de producció audiovisual.

ÚS DE LA IA GENERATIVA

Durant el desenvolupament d'aquest treball s'han utilitzat eines d'intel·ligència artificial generativa com a suport en la redacció i revisió del document, així com per a la consulta de conceptes tècnics i millora de l'expressió escrita.

Aquestes eines no han estat utilitzades per generar resultats experimentals ni implementacions directes del sistema, i totes les decisions tècniques i de disseny han estat preses de manera pròpia.

AGRAÏMENTS

Vull expressar el meu agraïment al meu tutor, Lluís Gomez, pel seu suport, orientació i disponibilitat al llarg del desenvolupament d'aquest TFG.

També vull agrair a 3Cat i a l'equip de professionals que m'ha acompanyat durant el projecte per la seva col·laboració, la informació proporcionada i la seva disponibilitat per compartir coneixements i experiències relacionades amb el procés de producció i adaptació de continguts audiovisuals.

Finalment, agraeixo a la Universitat l'oportunitat de desenvolupar aquest projecte i així acabar aquesta etapa acadèmica.

REFERÈNCIES

- [1] PySceneDetect Developers, *PySceneDetect Documentation*, <https://www.scenesdetect.com/>
- [2] OpenCV, *Histogram Comparison Tutorial*, https://docs.opencv.org/3.4/d8/dc8/tutorial_histogram_comparison.html
- [3] MathWorks, *Understanding Color Spaces and Color Space Conversion*, <https://www.mathworks.com/help/images/understanding-color-spaces-and-color-space-conversion.html>
- [4] J. Kukal, *Methods for Shot Boundary Detection*, Kybernetika, 1998, https://dml.cz/bitstream/handle/10338.dmlcz/135216/Kybernetika_34-1998-4_2.pdf
- [5] Y.-S. Chen et al., *Deep Auto-Reframing*, ACM Transactions on Graphics, 2018, <https://dl.acm.org/doi/10.1145/3231598>

- [6] J. Wang, J. Li, S. Lu, *Automatic Video Reframing via Content-Aware Retargeting*, arXiv:2101.04051, <https://arxiv.org/abs/2101.04051>
- [7] K. Stanuch, *Autocrop Vertical*, GitHub Repository, <https://github.com/kamilstanuch/Autocrop-vertical>
- [8] Universitat d'Alacant, *Tècniques de video retargeting i reenquadrament automàtic*, <https://rua.ua.es/server/api/core/bitstreams/53cc96e1-81bf-46e6-82ed-115f83289f7a/content>
- [9] LearnOpenCV, *Face Detection using OpenCV, Dlib and Deep Learning*, <https://learnopencv.com/face-detection-opencv-dlib-and-deep-learning-c-python/>
- [10] W. Wu, H. Peng, S. Yu, *YuNet: A Tiny Millisecond-Level Face Detector*, Machine Intelligence Research, vol. 20, no. 5, pp. 656–665, 2023, <https://link.springer.com/article/10.1007/s11633-023-1423-y>
- [11] J. Deng, J. Guo, N. Xue, S. Zafeiriou, *RetinaFace: Single-shot Multi-level Face Localisation in the Wild*, CVPR 2020, <https://arxiv.org/abs/1905.00641>
- [12] N. Wojke, A. Bewley, D. Paulus, *Simple Online and Realtime Tracking with a Deep Association Metric*, arXiv:1703.07402, <https://arxiv.org/abs/1703.07402>
- [13] LearnOpenCV, *Understanding DeepSORT*, <https://learnopencv.com/understanding-multiple-object-tracking-using-deepsort/>
- [14] MDPI, *Multiple Object Tracking Review*, Applied Sciences, 2022, <https://www.mdpi.com/2076-3417/12/11/5568>
- [15] J. Deng et al., *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, CVPR 2019, <https://arxiv.org/abs/1801.07698>
- [16] Ruhyadi, *Vision-FR*, GitHub Repository, <https://github.com/ruhyadi/vision-fr>
- [17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, CVPR 2016, <https://arxiv.org/abs/1506.02640>
- [18] OpenCV, *OpenCV 4.x Documentation*, <https://docs.opencv.org/4.x/>

APÈNDIX

A.1 Format de sortida JSON

El sistema desenvolupat exporta els resultats de l'anàlisi en un fitxer JSON. Aquest fitxer conté la informació necessària per descriure el reenquadrament vertical generat, incloent les dades generals del vídeo, la configuració utilitzada, els intervals de plans detectats i les regions d'interès calculades per a cada rang de fotogrames.

Aquesta sortida permet separar la fase d'anàlisi de la fase de generació del vídeo final. D'aquesta manera, les regions d'interès obtingudes poden ser reutilitzades posteriorment per altres eines de processament audiovisual sense haver de repetir tot el procés d'anàlisi.

Per reduir la mida del fitxer, les ROIs consecutives que tenen la mateixa posició i dimensions s'agrupen en intervals de frames. La Figura 9 mostra un exemple visual simplificat de l'estructura generada pel sistema.

```
"videoAnalysis": {
  "videoFilePath": "D:\\VIDEOS VERTICALS\\TN",
  "fps": 25.001980825608115,
  "totalFrames": 6311,
  "clipStartFrame": 1475,
  "clipEndFrame": 4850,
  "sceneCount": 36,
  "config": {
    "faceDetectorType": "YuNet",
    "faceModelFile": "..\\TESTOpenCVUtils\\",
    "faceConfigFile": "",
    "trackFaceFile": "..\\TESTOpenCVUtils\\",
    "personFile": "..\\TESTOpenCVUtils\\Mc",
    "roiWidth": 608,
    "roiHeight": 1080,
    "maxFaces": 1,
    "smoothAlpha": 0.0,
    "zoomEnabled": 0,
    "roiTransitionMode": "SmoothLimited",
    "detectPersonsEnabled": 0,
    "personMinAspectRatio": 1.799999952316,
    "personMinAreaRatio": 0.07999999821186,
    "clipStartSeconds": 59.0,
    "clipEndSeconds": 194.0
  },
  "scenes": [
    {
      "startFrame": 1475,
      "endFrame": 1493,
      "frames": [
        {
          "startFrame": 1475,
          "endFrame": 1487,
          "rois": [
            {
              "x": 656,
              "y": 0,
              "width": 608,
              "height": 1080
            }
          ]
        }
      ]
    }
  ]
},
```

Fig. 9: Exemple visual del fitxer JSON generat pel sistema.

En aquest exemple, el camp `config` descriu els paràmetres utilitzats durant l'anàlisi, com el detector facial emprat, la mida de la ROI vertical, el nombre màxim de cares considerades i el rang temporal processat. El camp `scenes` conté els diferents segments detectats dins del vídeo, i per a cada segment s'indiquen els intervals de frames i les ROIs associades.

El fitxer JSON complet corresponent a aquest exemple es pot consultar mitjançant l'enllaç següent:

Fitxer JSON

A.2 Vídeos resultants en format vertical

Amb l'objectiu de complementar l'avaluació qualitativa presentada en el document, s'han generat diversos vídeos resultants en format vertical. Aquests exemples permeten observar el comportament temporal del sistema, incloent l'estabilitat del reenquadrament, el seguiment dels subjec-

tes principals, l'ús del frame complet i la resposta davant escenes amb moviment o amb detecció facial limitada.

La Taula 2 recull els enllaços als vídeos generats pel sistema.

Vídeo	Descripció	Enllaç
1	Dues cares	Vídeo 1
2	Frame complet en un fragment del programa matins	Vídeo 2
3	Contingut informatiu sobre os-sos	Vídeo 3
4	Observatori	Vídeo 4
5	Contingut informatiu sobre han-tavirus	Vídeo 5
6	Explicació d'un cas amb frame complet	Vídeo 6
7	Notícia sobre la Vespa	Vídeo 7
8	Música, amb limitació per falta de detecció del parlant	Vídeo 8

TAULA 2: ENLLAÇOS ALS VÍDEOS RESULTANTS EN FORMAT VERTICAL.