Visión por Computador

Tema 5. Reconocimiento de objetos

EII, GII. *Modesto Castrillón Santana*

Contenidos

- Introducción
- Detección de objetos
- Segmentación semántica
- Clasificación
- Evaluación experimental
- Vídeo
- Visión y lenguaje

Introducción

Tareas VC:

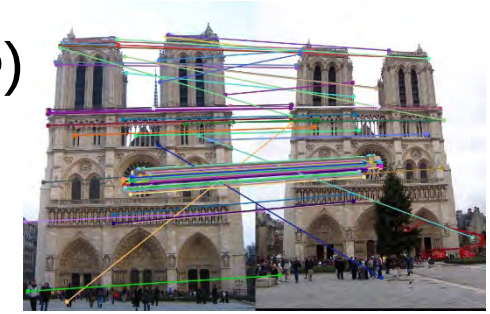
- Clasificación de imagen:
  - ¿Es un gato?
  - ¿Qué clase de escena es esta?

Introducción

Tareas VC:

- Clasificación de imagen:
  - ¿Es un gato?
  - ¿Qué clase de escena es esta?

- Reconocimiento de ejemplar (*instance*)
  - Localiza este edificio
  - Busca a esta persona

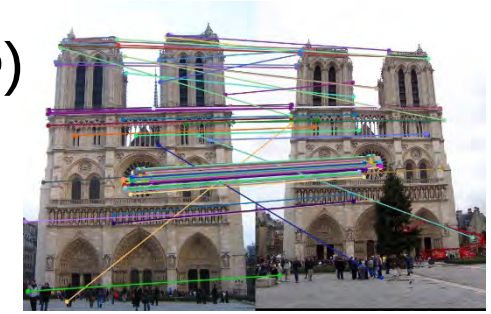


Introducción

Tareas VC:

- Clasificación de imagen:
  - ¿Es un gato?
  - ¿Qué clase de escena es esta?

- Reconocimiento de ejemplar (*instance*)
  - Localiza este edificio
  - Busca a esta persona



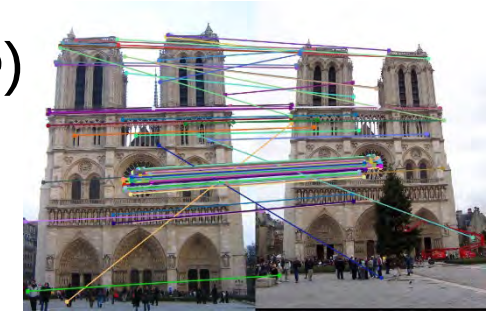
- Detección
  - Localiza coches, personas, etc.

Introducción

Tareas VC:

- Clasificación de imagen:
  - ¿Es un gato?
  - ¿Qué clase de escena es esta?

- Reconocimiento de ejemplar (*instance*)
  - Localiza este edificio
  - Busca a esta persona

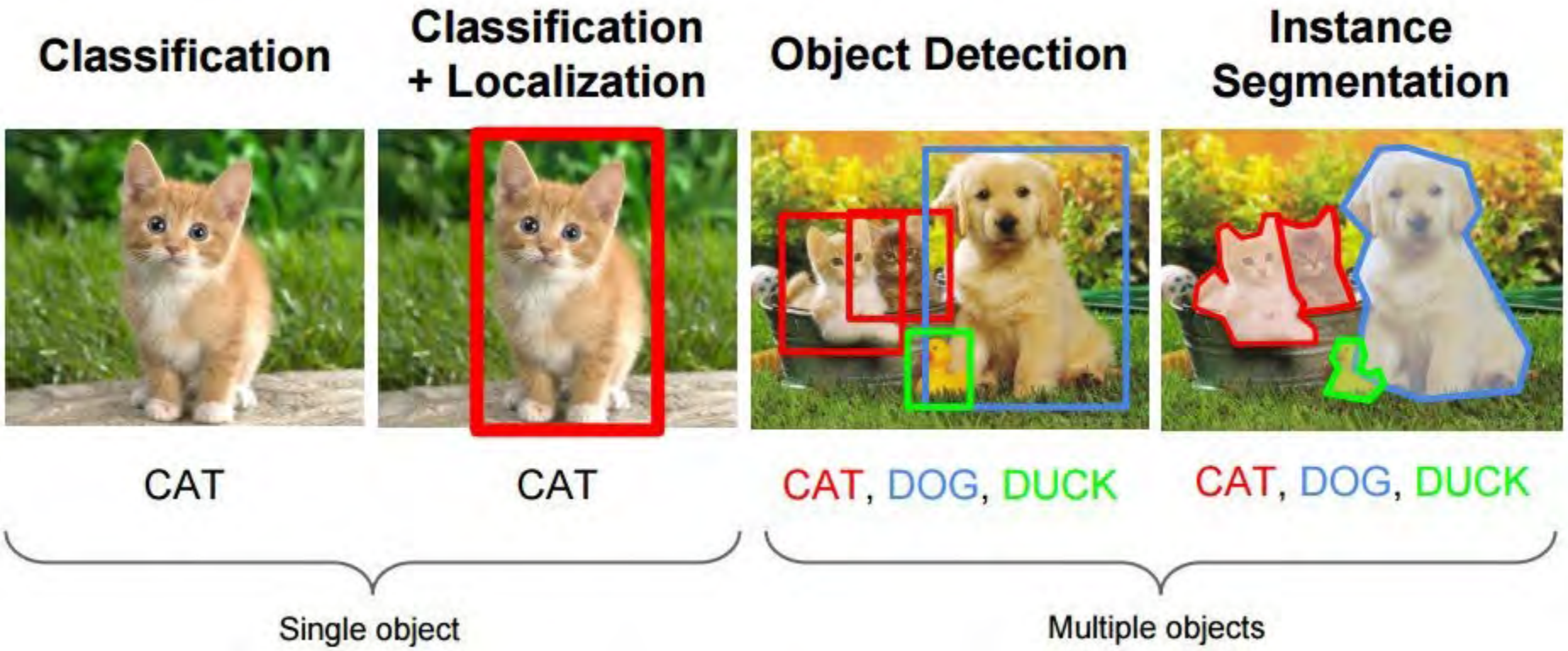


- Detección
  - Localiza coches, personas, etc.

- Comprensión de vídeo
  - ¿Qué está haciendo esta persona?

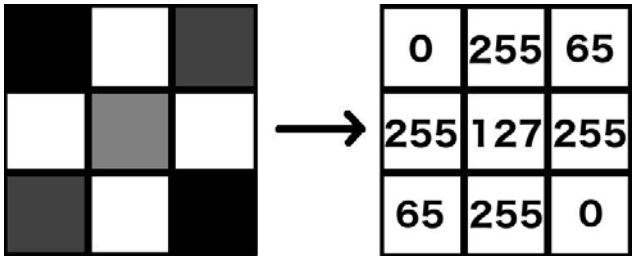
Introducción

Tareas VC:



Introducción

¿Dificilillo?



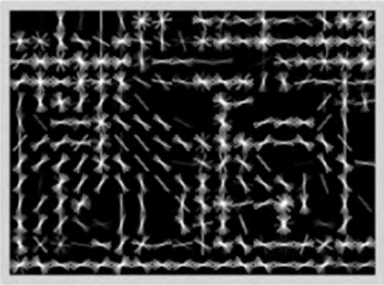
Introducción

Proceso general:

1. Especificar modelo
   - Estadístico
   - Articulado
   - Colección de ejemplares
   - Parametrizable
2. Generar hipótesis
   - Ventana deslizante
   - Puntos de interés
   - Regiones
3. Puntuar hipótesis
   - Confianza
   - Múltiples clasificadores
4. Resolver detecciones
   - Eliminación de no máximos (NMS)
   - Contexto

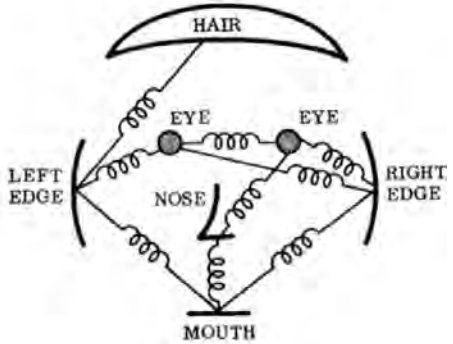
Introducción

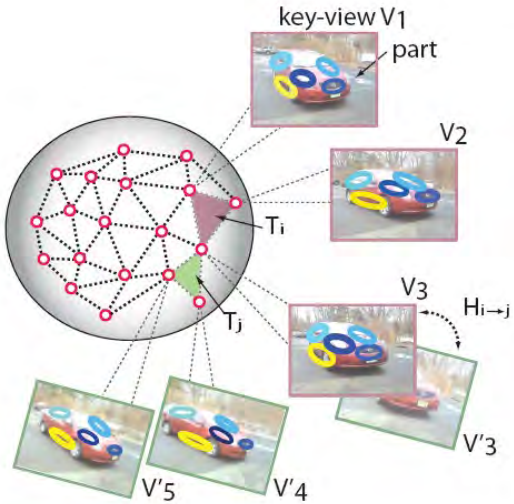
1. Especificar modelo
   - Estadístico
   - Articulado
   - Colección de ejemplares
   - Parametrizable



Modelo más complejo no necesariamente mejor

+ complejidad -> + costoso





Introducción

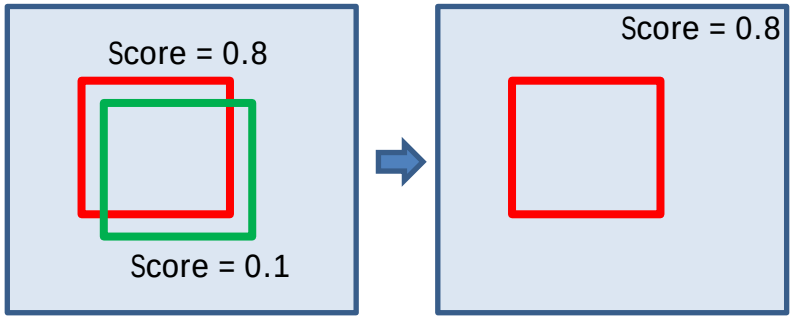
2. Generar hipótesis
   - Ventana deslizante
   - Puntos de interés
   - Regiones

B. Leibe et al.

Endres Hoiem

Introducción

3. Puntuar hipótesis
   - Confianza
   - Múltiples clasificadores
4. Resolver detecciones
   - Eliminación de no máximos (*NMS*)
   - Contexto



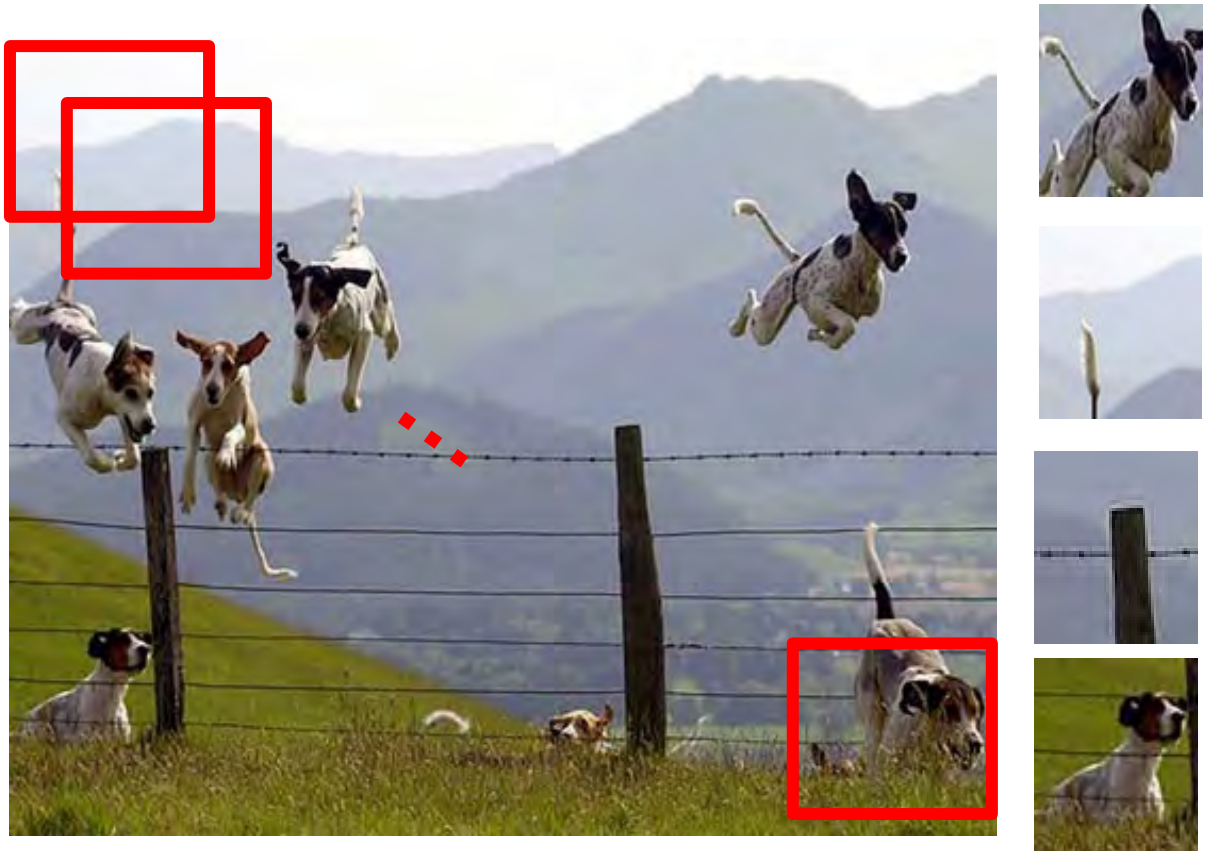
James Hays

(g) Car Detections: Local (h) Ped Detections: Local

Hoiem et al.

Detección

Localizar objeto en la imagen
Modelar el objeto



¿Ser o no ser?

James Hays

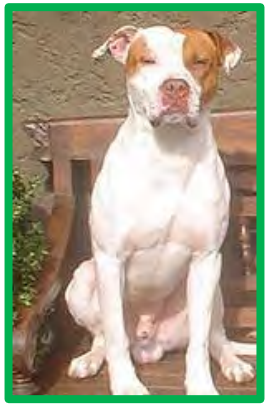
Detección

Comprobar multitud de ventanas/zonas y escalas
Selección de características
Medidas de encaje/similitud
Cambios de puntos de vista

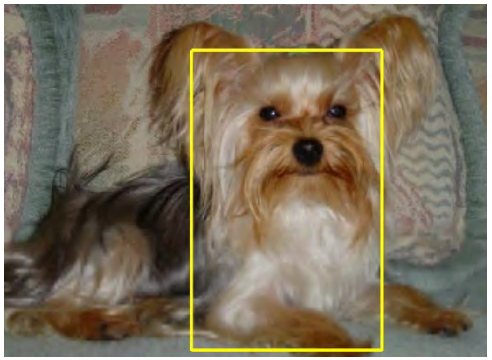
Detección

Modelar el objeto

TP



Mala localización



Similares

Diferentes

James Hays

Detección

Retos

Iluminación

Pose

Fondo

Oclusiones

Apariencia intraclase

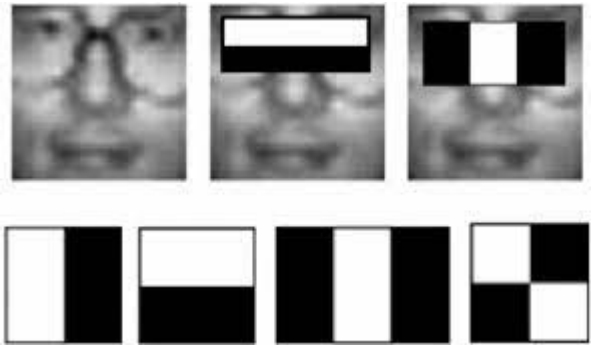
Punto de vista

K. Grauman, B. Leibe

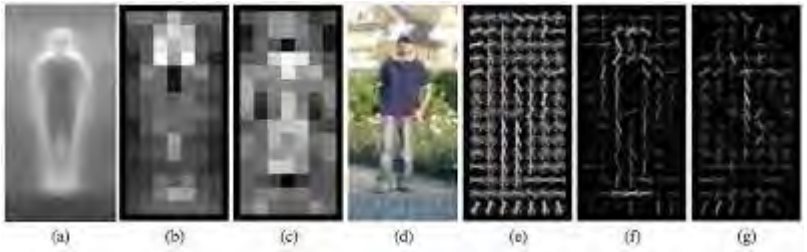
Detección

Enfoques previos a DL:

- Viola y Jones (caras)
  - *Boosting*
  - Árboles de decisión



- Dalal y Triggs (personas)
  - HOG
  - SVM

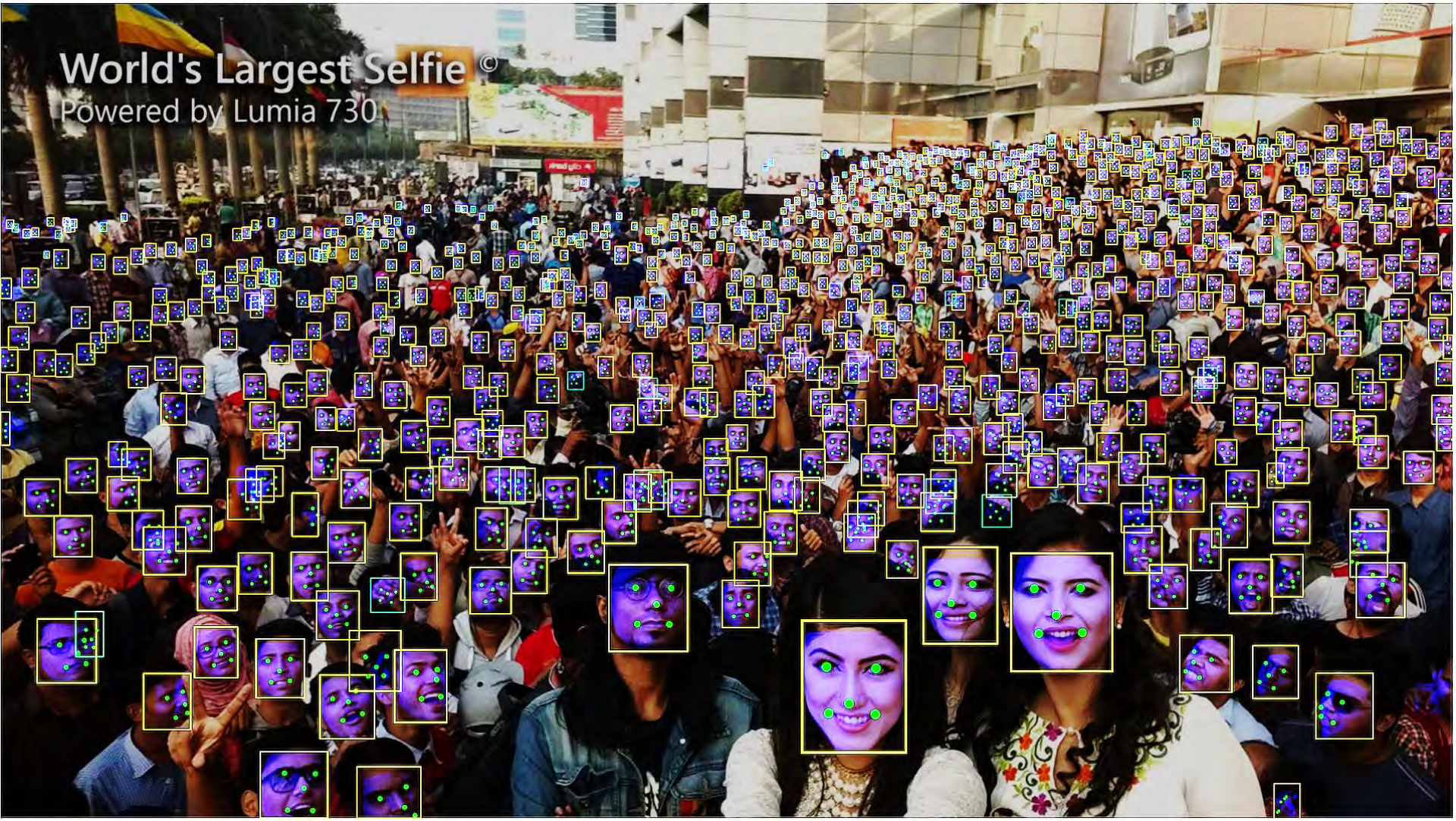


Aplicado tb a otras entidades

World's Largest Selfie ©
Powered by Lumia 730

P. Viola, M. Jones, M. *Rapid object detection using a boosted cascade of simple features* CVPR 2001

V. Kazemi and J. Sullivan. *One Millisecond Face Alignment with an Ensemble of Regression Trees*. CVPR14



J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. 2019. *RetinaFace: Single-stageDense Face Localisation in the Wild.* 2019

V. Albiero, X. Chen, X. Yin, G. Pang, T. Hassner, *img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation*, 2020

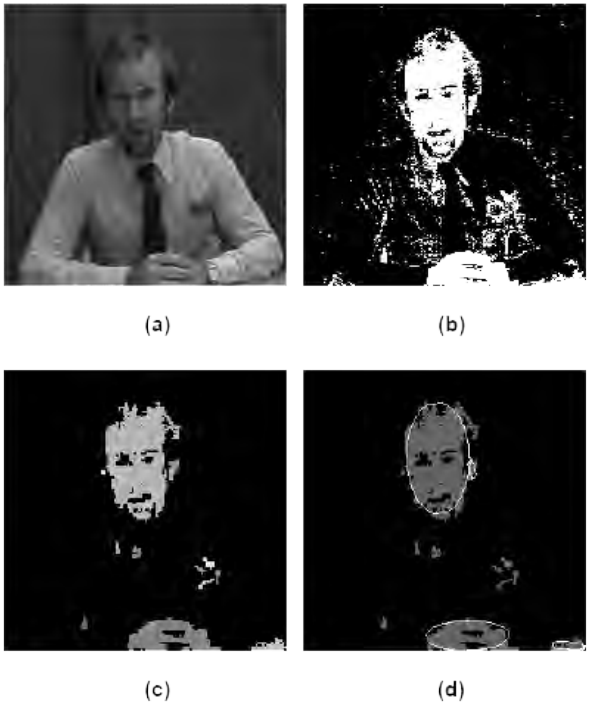
Caras

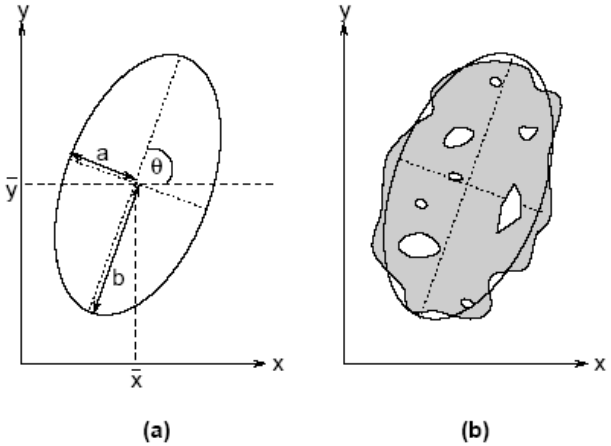
s. XX: heurísticas como color y movimiento, restricciones

Sobottka et al.

Color zonas similares a la piel humana (*inRange*)

Componentes conectadas con **restricciones geométricas**





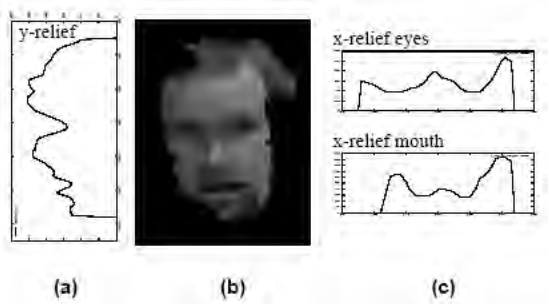
K. Sobottka,I. Pitas. A novel method for automatic face segmentation, face feature extraction and tracking. Signal Processing: Image Communication, 1998.

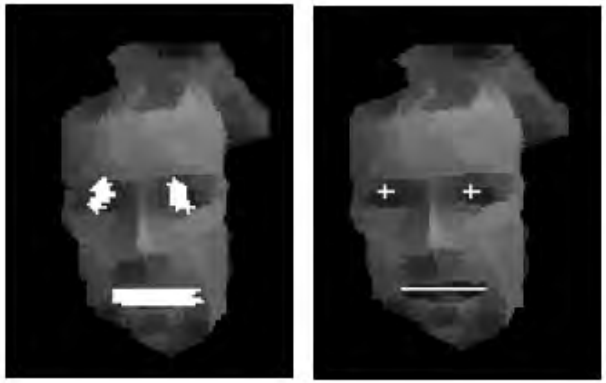
Caras

s. XX: heurísticas como color y movimiento, restricciones

Sobottka et al.

Asumiendo **cara frontal**, estima **ojos** por proyecciones



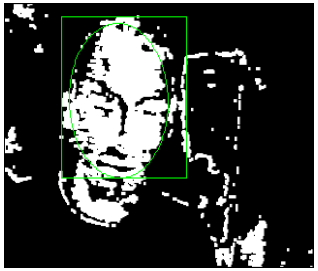


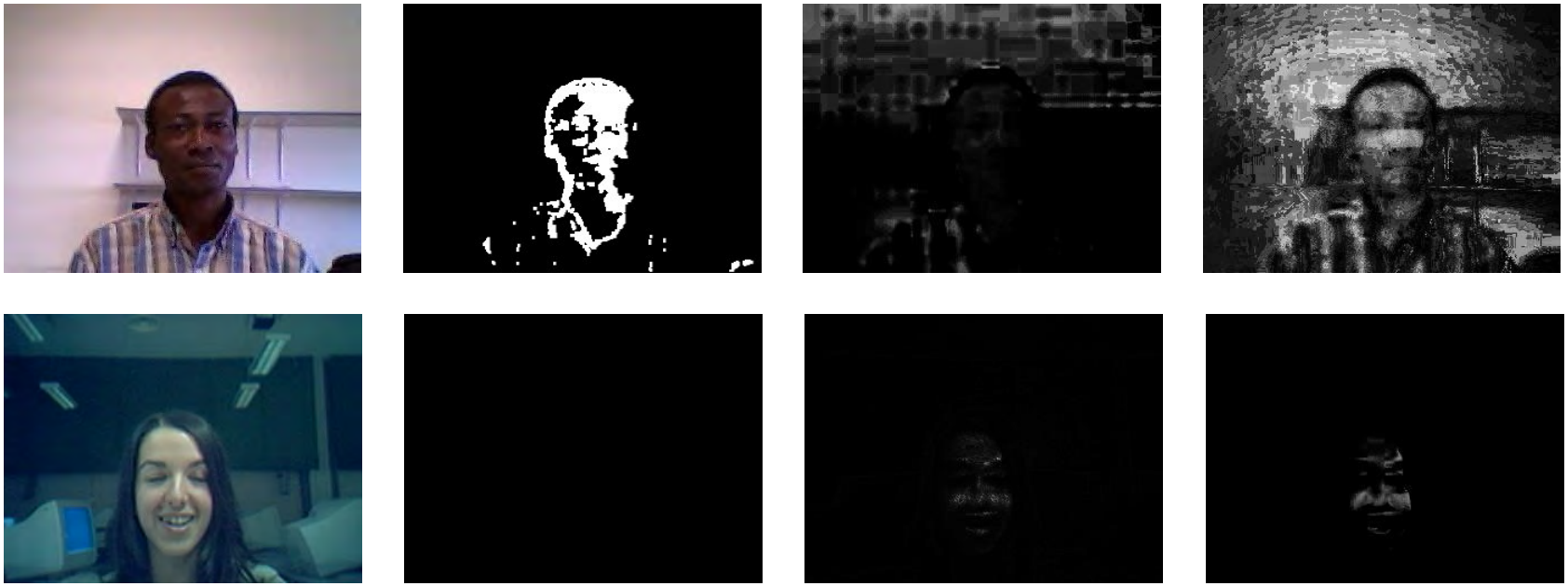
K. Sobottka,I. Pitas. A novel method for automatic face segmentation, face feature extraction and tracking. Signal Processing: Image Communication, 1998.

Caras

s. XX: color y movimiento, restricciones

Técnicas eficientes
No robustas





Esquema simple

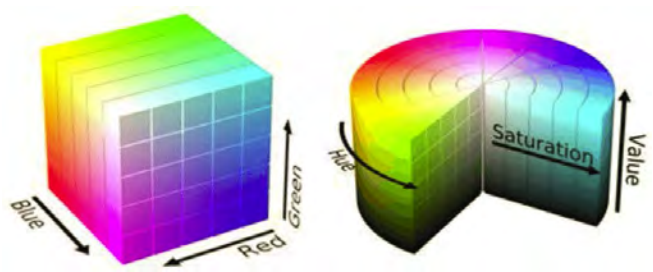
Jones and Rehg [1]

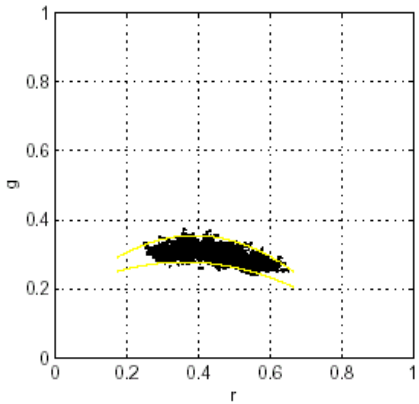
Kruppa [2]

Caras

s. XX: heurísticas como color y movimiento, restricciones

No robustas
Sin modelo general para el color piel

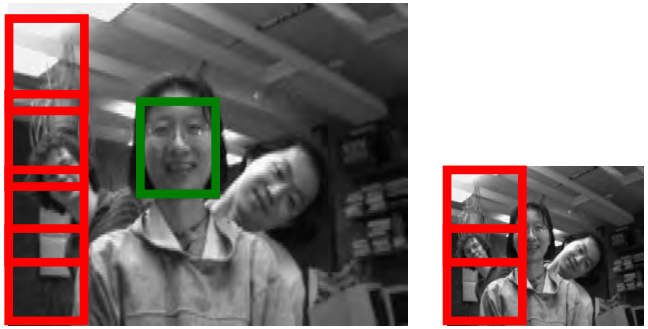




M. Störring et al.. Physics-based modelling of human skin colour under mixed illuminants. Robotics and Autonomous Systems, 2001.

Caras

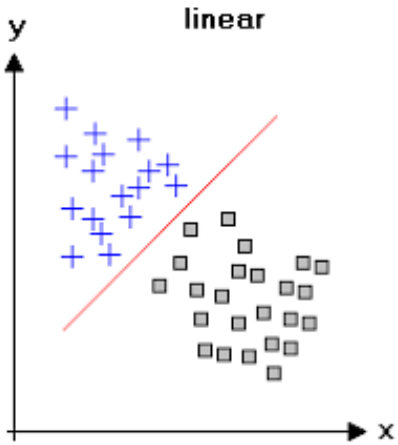
Aprendizaje automático
No heurísticas
Ventana deslizante
Coste temporal del clasificador



Caras

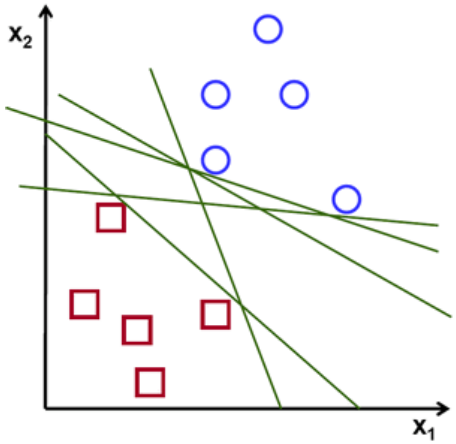
Clasificar cada ventana

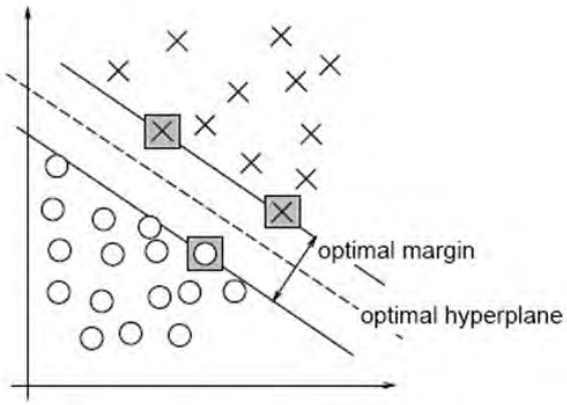
$$x \in \Re^n \,, \quad y \in \{\pm 1\}$$

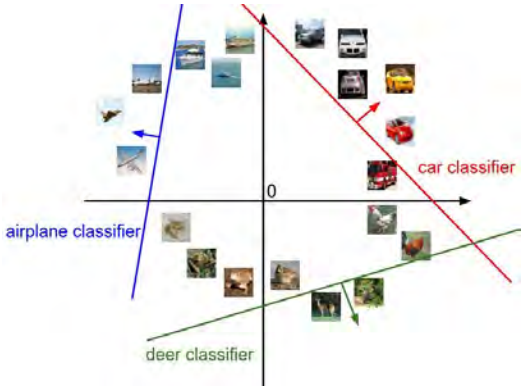


Caras

Clasificar cada ventana







Caras

Rowley et al.

Supervisado

Modela clases **cara** y **no cara** en zonas próximas

Negativas con *bootstrap*

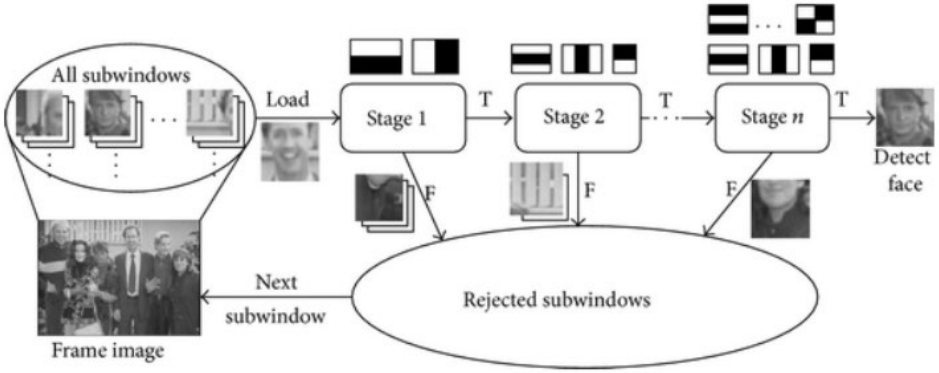
Define mejor la frontera entre ambas clases

Clasificador basado en redes neuronales

Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 1998.

Caras

Viola y Jones (>44k citas)
Hito en detección en "*tiempo real*"
Clasificador en cascada, clasificadores débiles
Reduce tiempo en zonas poco prometedoras
Características de cómputo rápido
Entrenamiento supervisado lento
Detección rápida
Imagen integral
Boosting



Cascade structure for Haar classifiers.

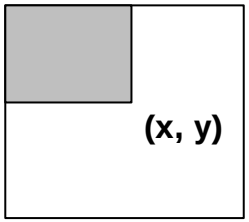
P. Viola and M. J. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In Computer Vision and Pattern Recognition, 2001.

Caras

Viola y Jones

Imagen integral

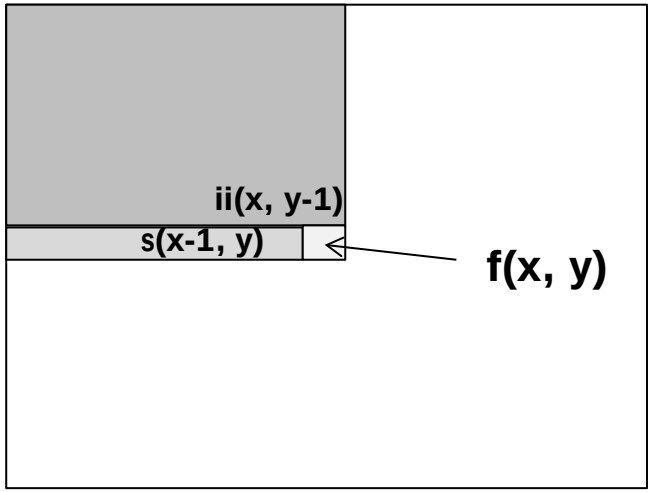
Valor del píxel, suma situados encima y a la izquierda



Cálculo en una pasada:

$$s(x,y) = s(x-1,y) + f(x,y)$$

$$ii(x,y) = ii(x,y-1) + s(x-1,y) + f(x,y)$$



Caras

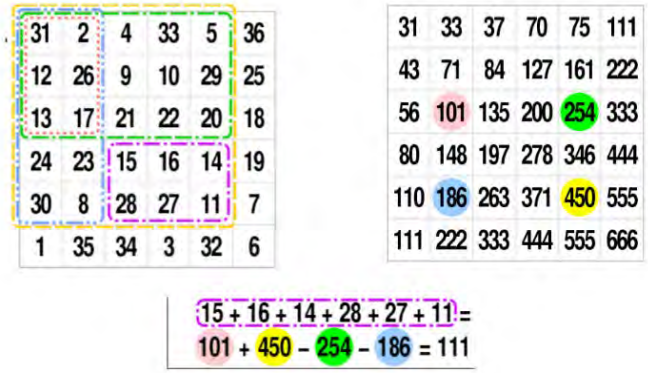
Viola y Jones

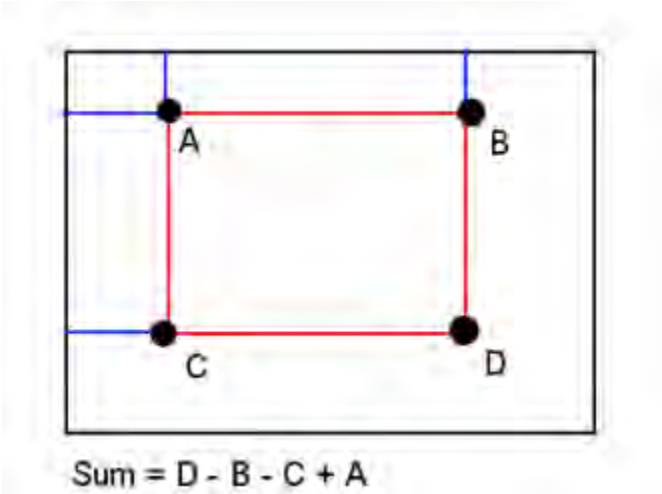
Imagen integral

Cálculo de suma en zona rectangular

$$R = D - B - C + A$$

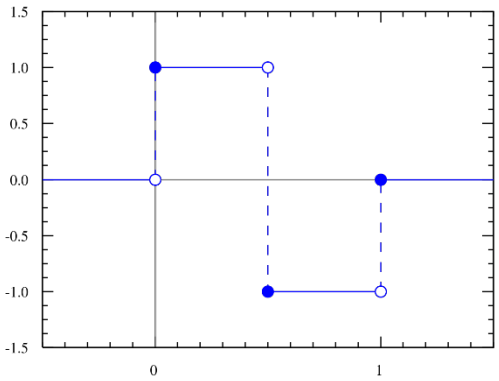
Tres sumas



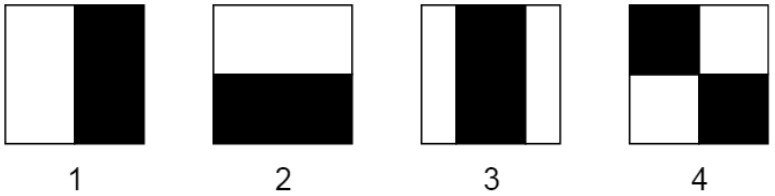


Caras

Viola y Jones
Características
Basadas en Haar *wavelet*



Diferencias de sumas de intensidad en zonas rectangulares
Zonas blancas suman, zonas negras restan
Variaciones tamaño y posición



Caras

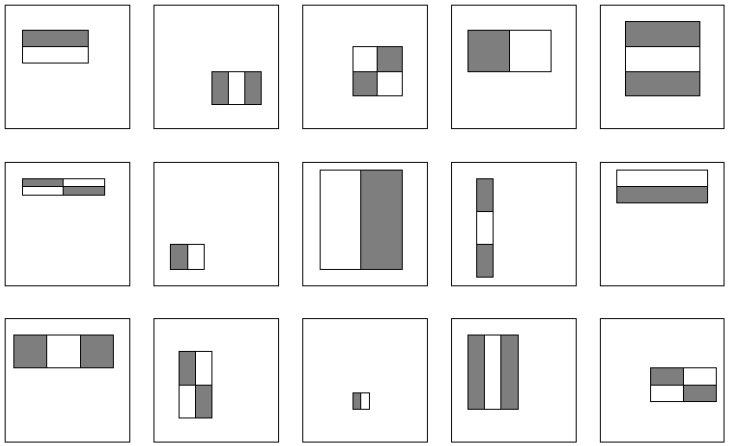
Viola y Jones

Características

Evaluando miles de combinaciones

Para imágenes 24x24 ~160000 combinaciones

¿Qué característica individual acierta más con las muestras?



Caras

Viola y Jones

Características

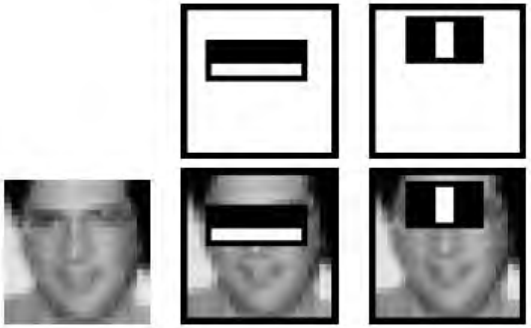
Construir un clasificador con un subconjunto

Clasificador débil con mejores características

¿Cuáles funcionan mejor para mi conjunto de datos?

Aplica *boosting*

Iterar modificando peso de muestras mal clasificadas

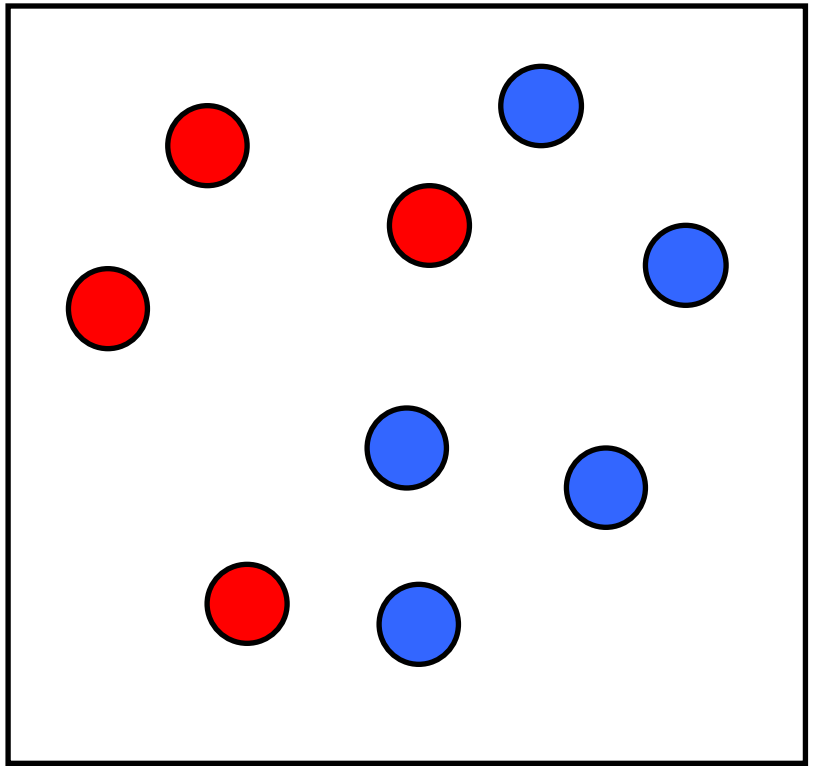


Caras

Viola y Jones

Boosting

Inicio, peso homogéneo para todas las muestras



Paul Viola

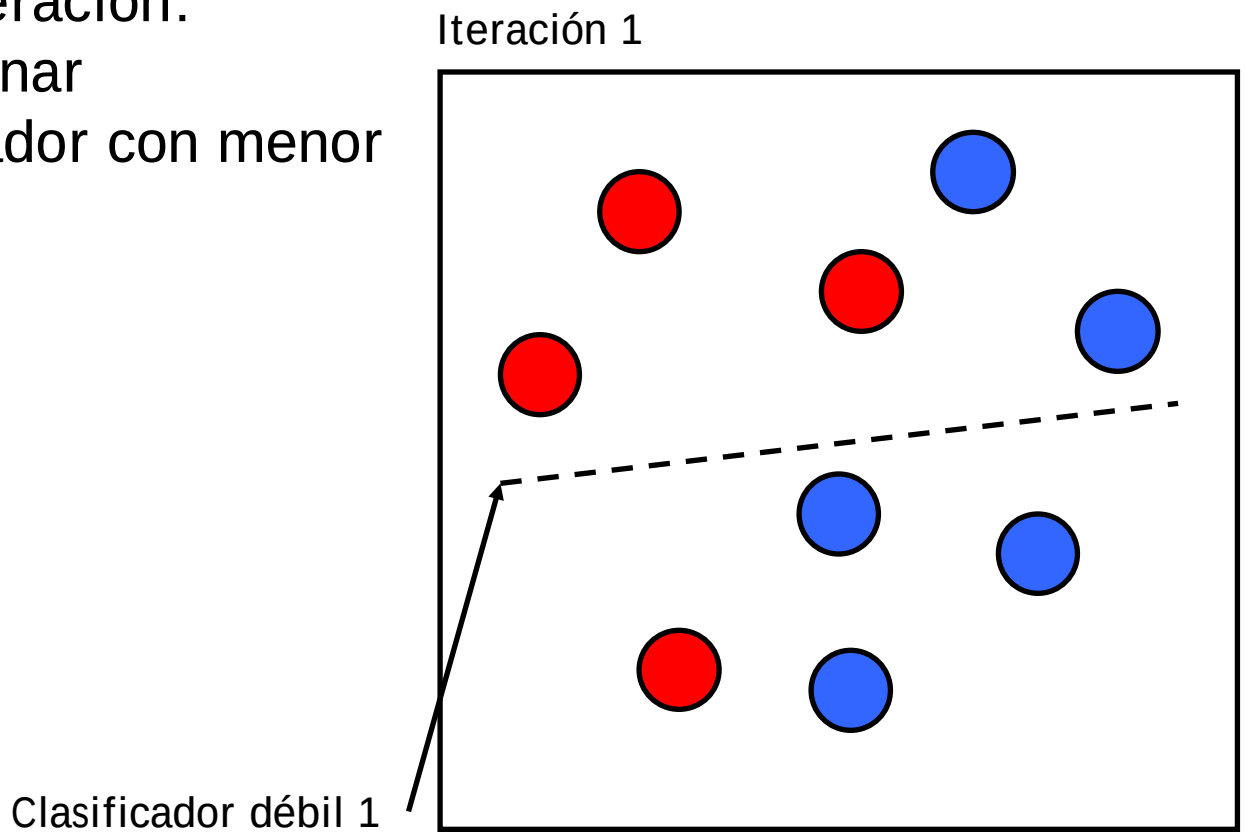
Caras

Viola y Jones

Boosting

En cada iteración:

- Determinar clasificador con menor error



Paul Viola

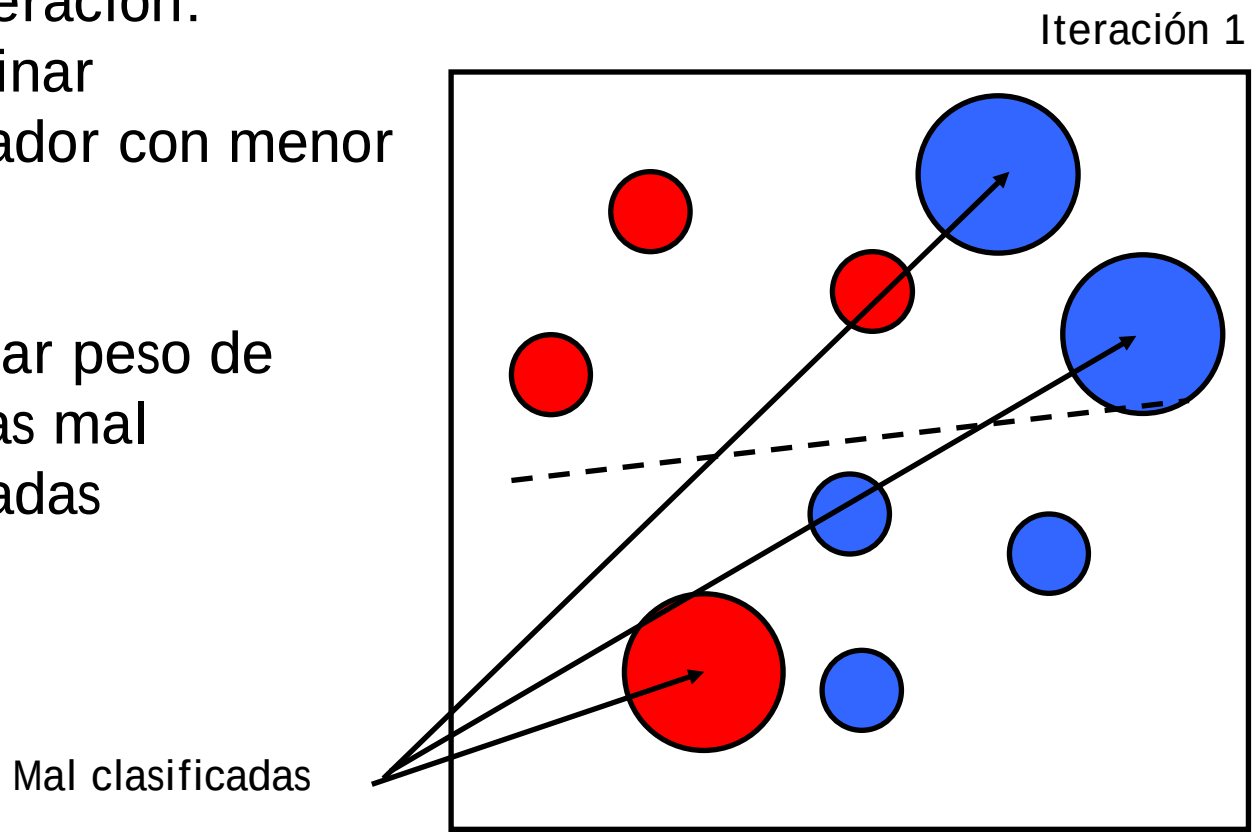
Caras

Viola y Jones

Boosting

En cada iteración:

- Determinar clasificador con menor error
- Aumentar peso de muestras mal clasificadas



Paul Viola

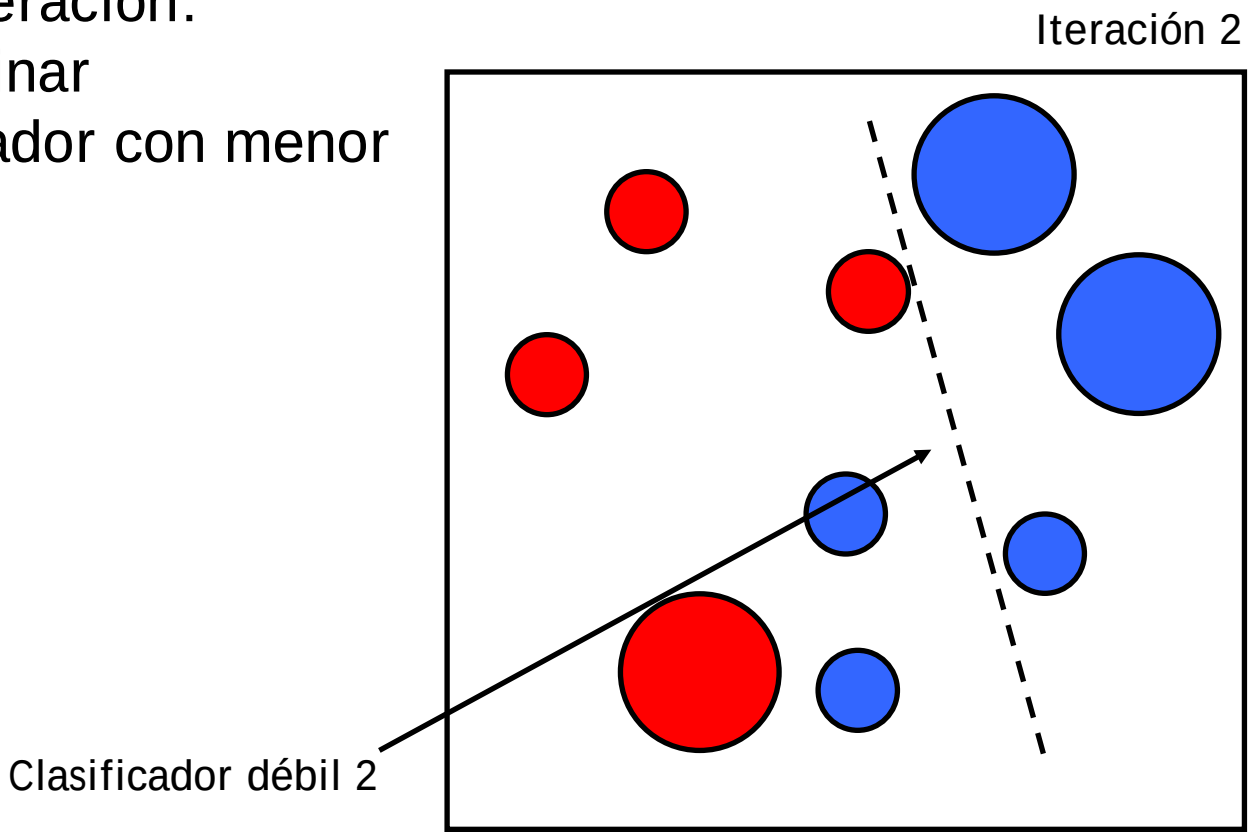
Caras

Viola y Jones

Boosting

En cada iteración:

- Determinar clasificador con menor error



Paul Viola

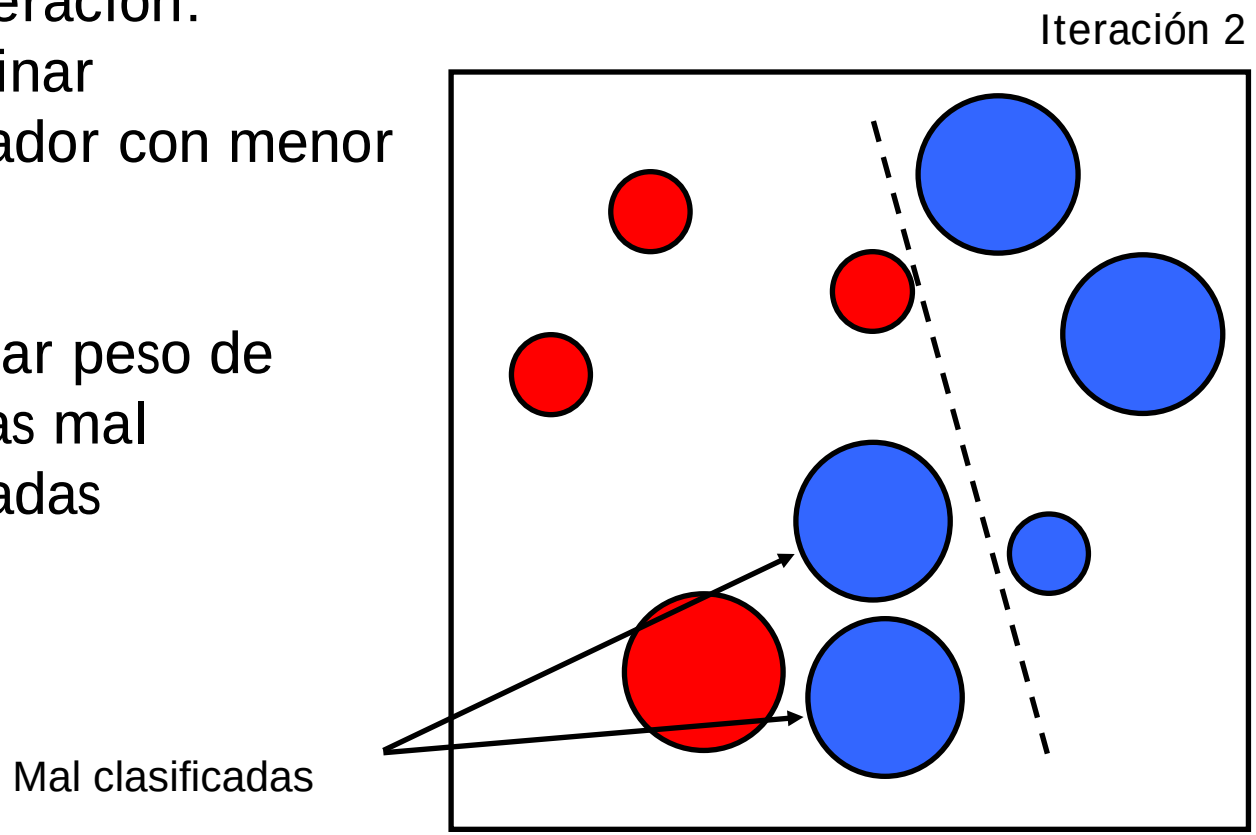
Caras

Viola y Jones

Boosting

En cada iteración:

- Determinar clasificador con menor error
- Aumentar peso de muestras mal clasificadas



Paul Viola

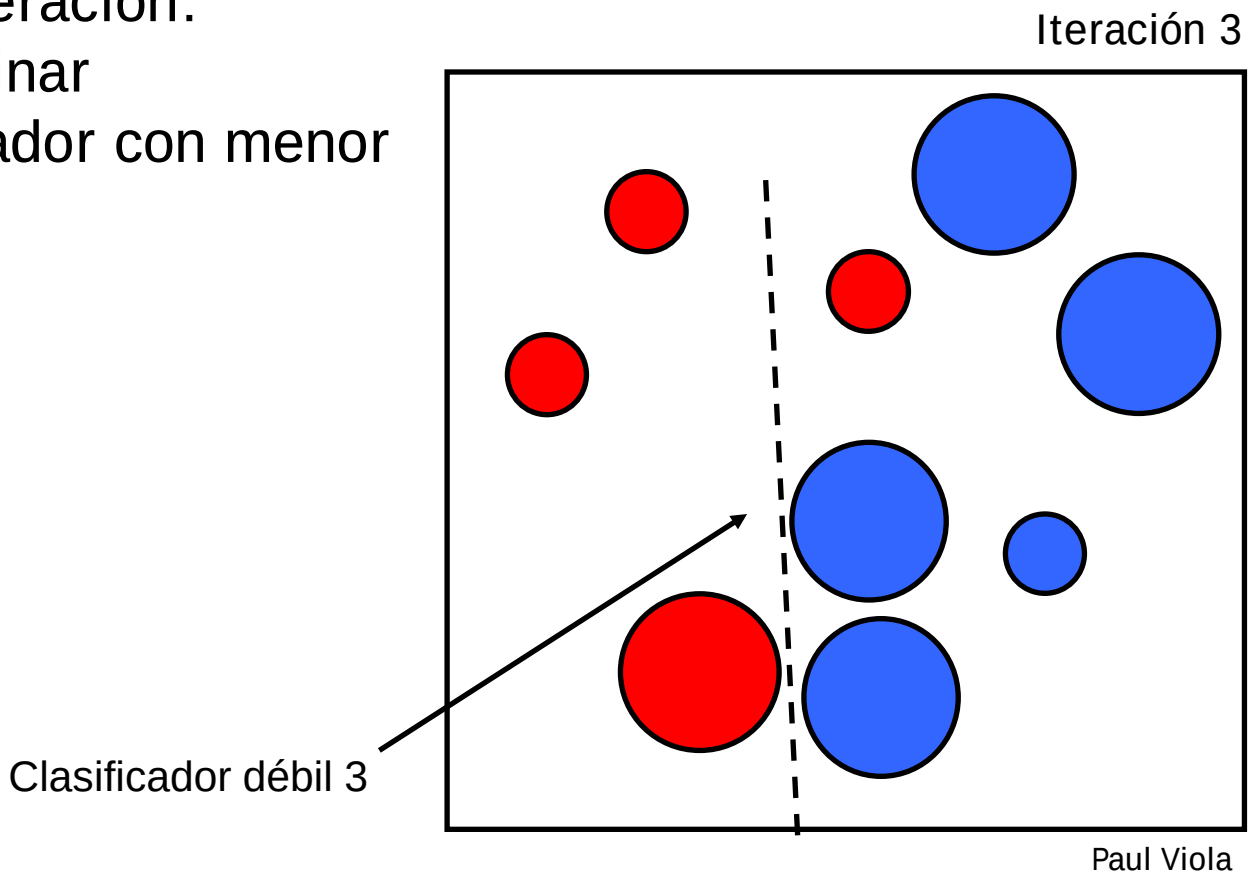
Caras

Viola y Jones

Boosting

En cada iteración:

- Determinar clasificador con menor error
- ...



Caras

Viola y Jones

Boosting

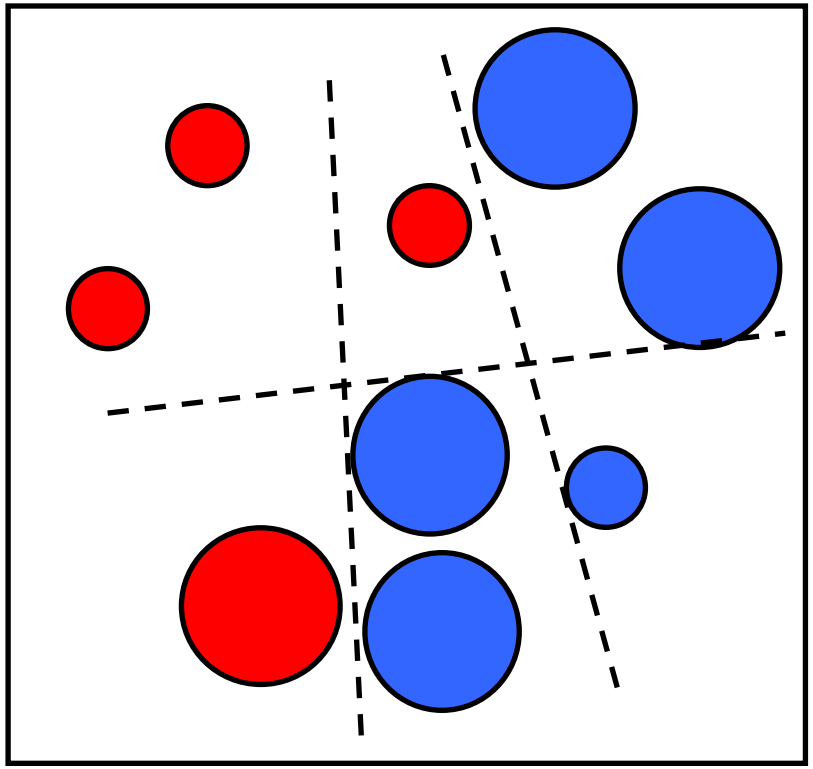
Tras n iteraciones

Combinación lineal (hay variantes) de clasificadores débiles (h_i)

$$h(x) = sgn\left(\sum_{j=1}^{M} \propto_j h_j(x)\right)$$

Cada clasificador débil:

$$h_j(x) = \begin{cases} -s_j & si \quad f_j < \theta_j \\ s_j & en \ \ otro \ caso \end{cases}$$



Paul Viola

Caras

Viola y Jones

Boosting vs. clasificador fuerte (p.e. SVM)

Ventajas

Integra entrenamiento y selección de características

Evaluación rápida

Flexibilidad en elección de esquema de *boosting*

Desventajas

Entrenamiento lento

Necesita muchas muestras

A menudo no tan bueno como SVM (multiclase)

Caras

Viola y Jones

Entrenamiento

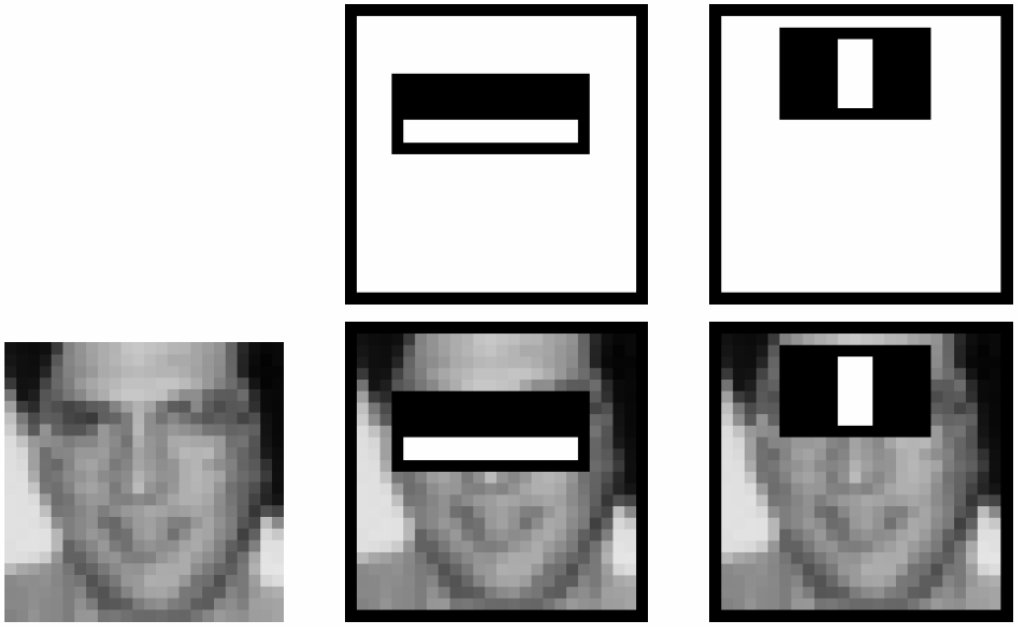
Realiza *boosting* añadiendo clasificadores débiles

Buscando cierta tasa de TPR y FPR en cada etapa

Muestras negativas >> positivas

Caras frontales

100% *recall* y 50% FPR



Caras

Viola y Jones

No basta, FPR alta

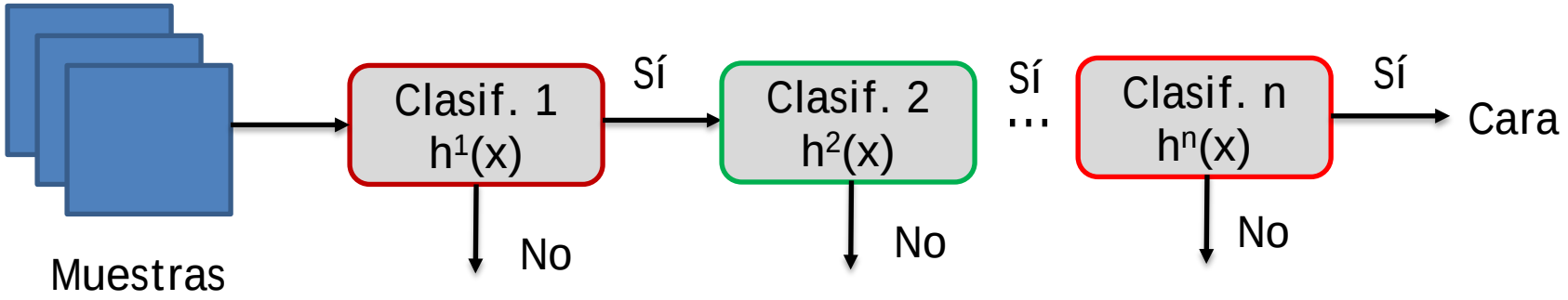
Ventana deslizante -> muchas hipótesis

No hay tantas caras

¿Cómo reducir el tiempo de zonas *fáciles*?

Atención, clasificador en cascada

Alcanzar en cada etapa una cierta tasa de FPR



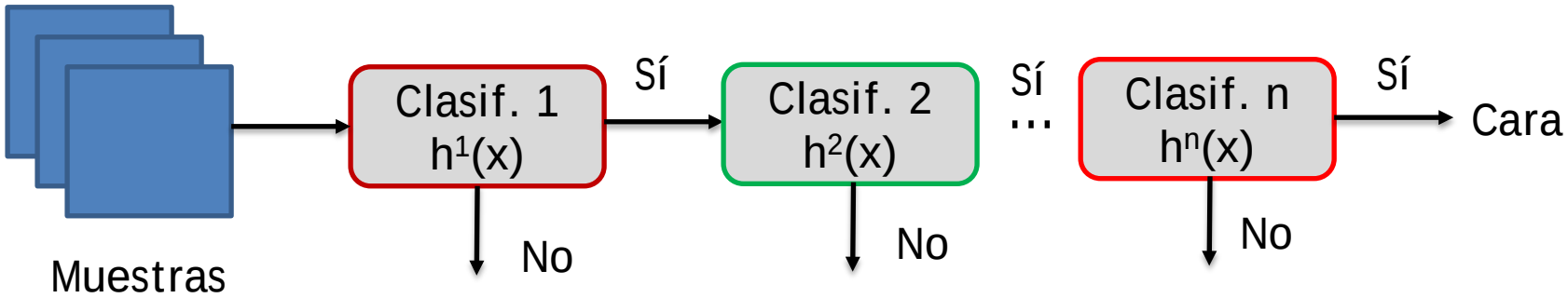
Caras

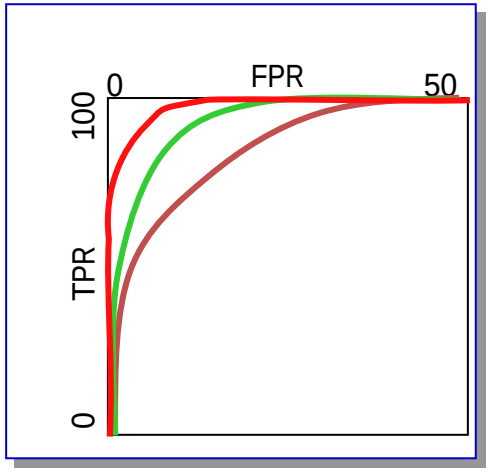
Viola y Jones

Clasificador en cascada

Siguiente nivel entrena con muestras aceptadas por anterior

Cada nivel clasificador más complejo





Caras

Viola y Jones

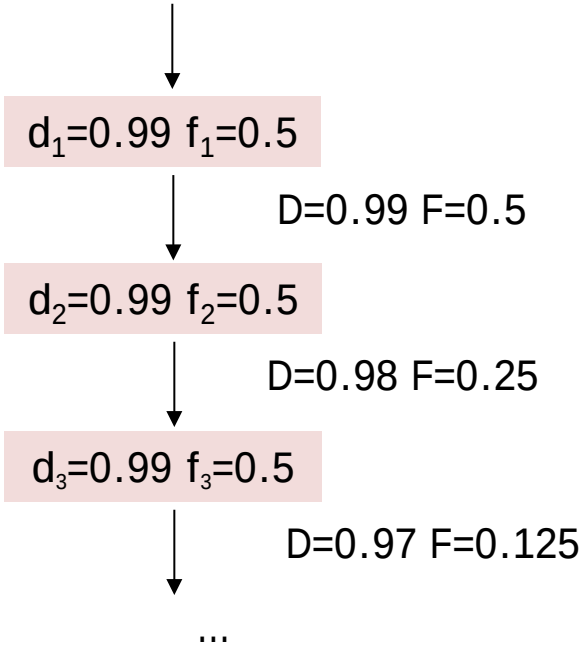
Clasificador en cascada

Siguiente nivel entrena con muestras aceptadas por anterior

Cada nivel clasificador más complejo

Ratios resultantes de detecciones, D, y de falsas detecciones, F, es la combinación de los de cada etapa:

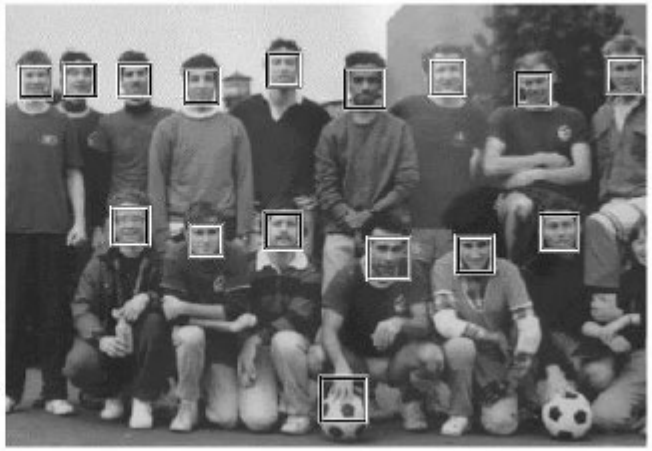
Detector de **20 etapas**, etapas que rechazan un 50% de patrones no faciales, y sólo rechaza 0.1% de caras. Rendimiento general: tasa de **detección de 0.999^{20}**, tasa de **falsas detecciones de 0.5^{20}**



Caras

Viola y Jones

Resultados



Detector \ False detections	10	31	50	65	78	95	167
Viola-Jones	76.1%	88.4%	91.4%	92.0%	92.1%	92.9%	93.9%
Viola-Jones (voting)	81.1%	89.7%	92.1%	93.1%	93.1%	93.2 %	93.7%
Rowley-Baluja-Kanade	83.2%	86.0%	-	-	-	89.2%	90.1%
Schneiderman-Kanade	-	-	-	94.4%	-	-	-
Roth-Yang-Ahuja	-	-	-	-	(94.8%)	-	-

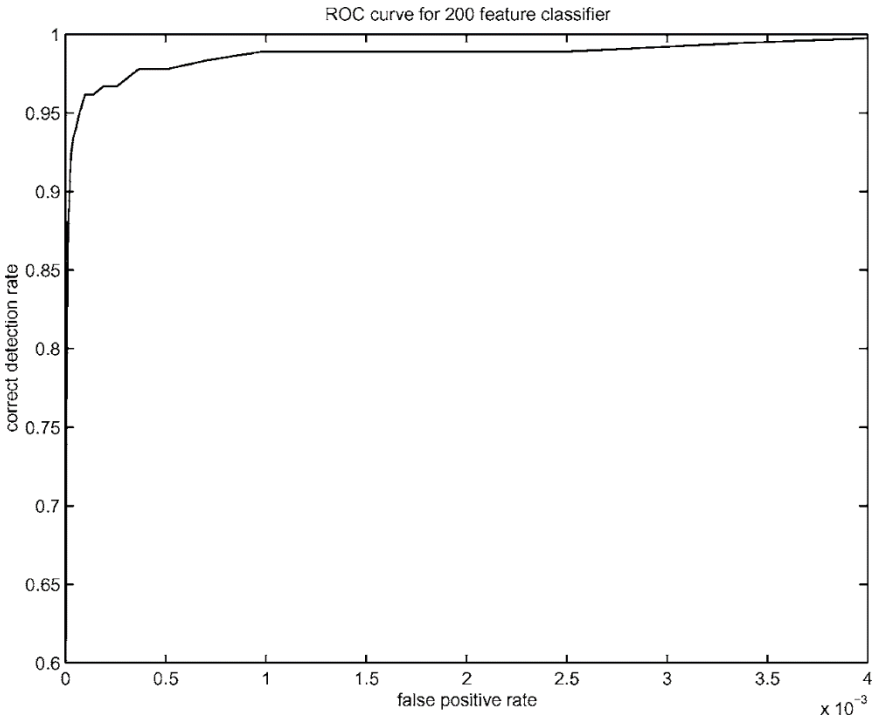
MIT + CMU face dataset

Caras

Viola y Jones

Modelos disponibles en OpenCV desde 2002 (R. Leinhart)

Democratizó acceso a detectores faciales



Caras

Viola y Jones

No suficiente

Falsas detecciones

Caras

Viola y Jones

No suficiente

Insuficientes detecciones

No sólo hay caras frontales, las caras pueden estar rotadas:

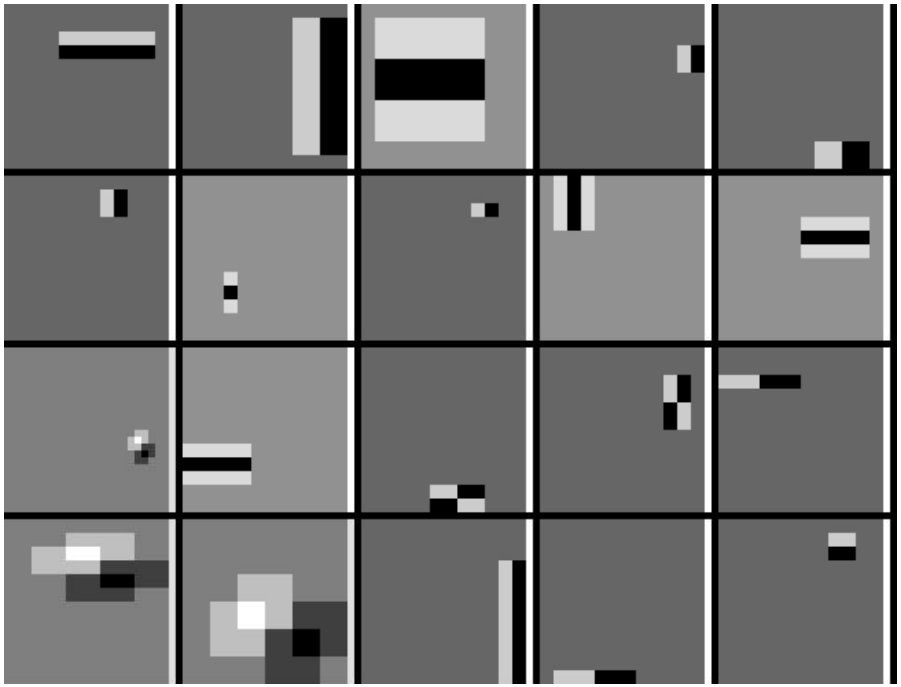
- Rotación fuera del plano (*out-of-plane*)
- Rotación en el plano (*in-plane*)
- Rotación arriba y abajo

Rowley ya proponía usar + clasificadores

Caras

Viola y Jones

Perfil



Caras

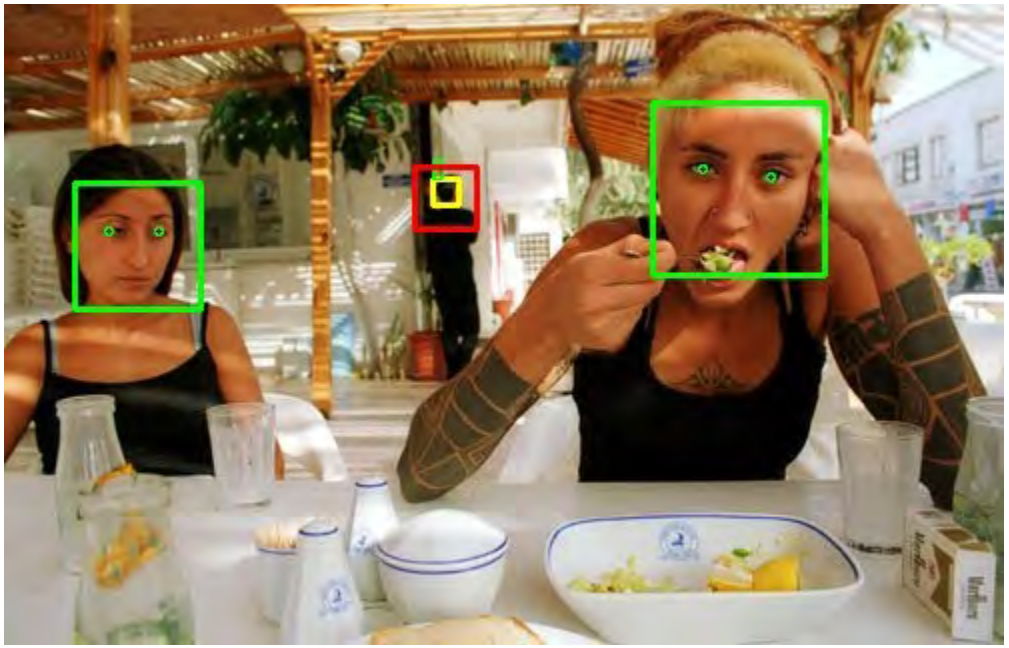
Viola y Jones

Cabeza y hombros

H. Kruppa, M. Castrillon-Santana, B. Schiele, Fast and Robust Face Finding via Local Context, PETS 2003

Caras

Viola y Jones



Caras

[Batallita]
Tesis sobre detección y seguimiento de caras
Dependencia del color piel
Visito el ETH en 2003 para conocer a Hannes Kruppa

Caras

[Batallita]
Desestimamos color
Viola Jones disponible en OpenCV desde 2002
Aprendemos a entrenar detectores
Incluidos en OpenCV desde 2010 a 2021

- The reference manual has been converted from HTML to LaTeX (by *James Bowman and Caroline Pantofaru*), so there is now:
  - opencv.pdf for reading offline
  - and the online up-to-date documentation (as the result of LaTeX→Sphinx→HTML conversion) available at
    - http://opencv.willowgarage.com/documentation/index.html
- **Samples, misc.:**
- Better eye detector has been contributed by *Shiqi Yu*, **see opencv/data/haarcascades/**[lefteye|righteye]*.xml
- sample LBP (Local Binary Pattern) cascade for the frontal face detection has been created by ***Maria Dimashova_**, see opencv/data/lbpcascades/lbpcascade*frontalface.xml
- Several high-quality body parts and facial feature detectors have been
  - *contributed by Modesto Castrillon-Santana*,
  - see opencv/data/haarcascades/haarcascade_mcs*.xml

Tutorial en pdf *(esp)*

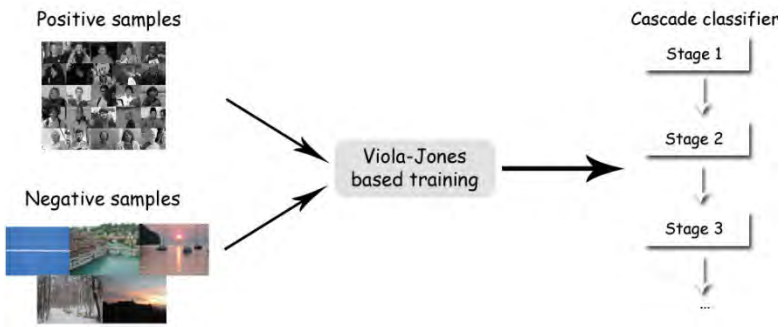
Tutorial haartraining (eng)

Nuestros Clasificadores para OpenCV

Rainer Lienhart and Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. ICIP2002

Caras

[Batallita]
Entrenando un clasificador VJ
Descargar binarios OpenCV



Dos conjuntos:

1. **Muestras positivas**:
   - ***opencv_createsamples*** genera entrenamiento en formato OpenCV
   - ***vec*** fichero resultante
   - Número de muestras positivas se indica con ***num***
   - ***info*** especifica fichero de texto con imágenes del patrón de interés
   - Muestras positivas, no necesariamente del mismo tamaño, ***opencv_createsamples*** escala en base a valores de ***w*** y ***h***

opencv_createsamples -info positives/trainpos.txt -vec data/positives.vec -num 5000 -w 22 -h 20

Caras

[Batallita]

Entrenando un clasificador VJ

Cada fila de trainpos.txt indica: 1) ruta del fichero, 2) número de objetos en la imagen y 3) sus contenedores

Contendedor: Coordenadas x e y de la esquina superior izquierda, seguidas del ancho y alto del contenedor.

Samples/subject01_centerlight.png 1 126 88 121 137

Samples/subject01_glasses.png 1 116 80 121 137

Caras

[Batallita]
Entrenando un clasificador VJ

Dos conjuntos:

2. **Muestras negativas**:
   - Fichero texto
   - Cada fila del fichero una imagen sin presencia del patrón
   - Requiere ruta absoluta
   - Imágenes grandes estilo fondo de escritorio

C:\Users\usuario\...\samples\000001.jpg

C:\Users\usuario\...\ \samples\01y_fenster.JPG

Caras

[Batallita]

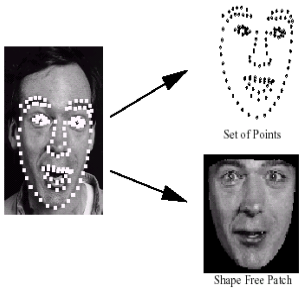
Código Python

```
import cv2 # OpenCV Library

# Carga clasificador
Cascade = cv2.CascadeClassifier('./Cascades/classifier.xml')

# captura de la cámara
video_capture = cv2.VideoCapture(0)
while True:
	# Obtiene fotograma
	ret, frame = video_capture.read()
	gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)

	# Detecta objetos y dibuja contenedores
	objects = Cascade.detectMultiScale(gray)
	for (x, y, w, h) in objects:
		container = cv2.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)

	# Muestra resultados
	cv2.imshow('Video', frame)
	if cv2.waitKey(1) & 0xFF == ord('q'):
		break

# Libera
video_capture.release()
cv2.destroyAllWindows()
```

Caras

***Active Shape/Appearance Models* (ASM/AAM)**

Modelo estadístico para la forma (contornos) y apariencia



Actualizado: **STASM**

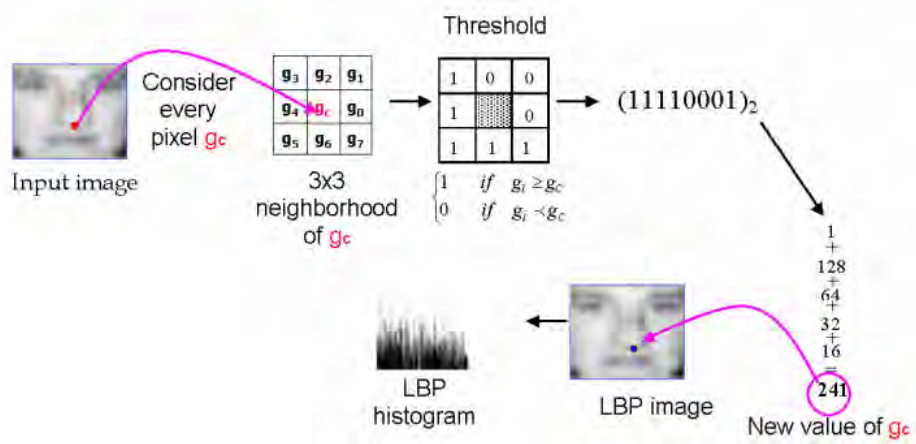
T. Cootes and C. Taylor. Statistical Models of Appearance for Computer Vision, University of Manchester. Report University of Manchester, 2000.
S. Milborrow and F. Nicolls. Active Shape Models with SIFT Descriptors and MARS. VISAPP, 2014

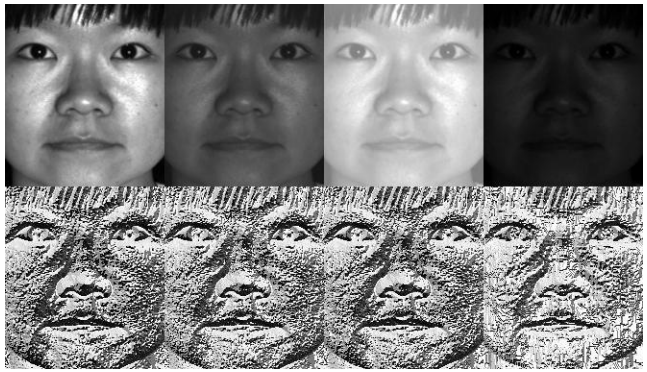
Caras

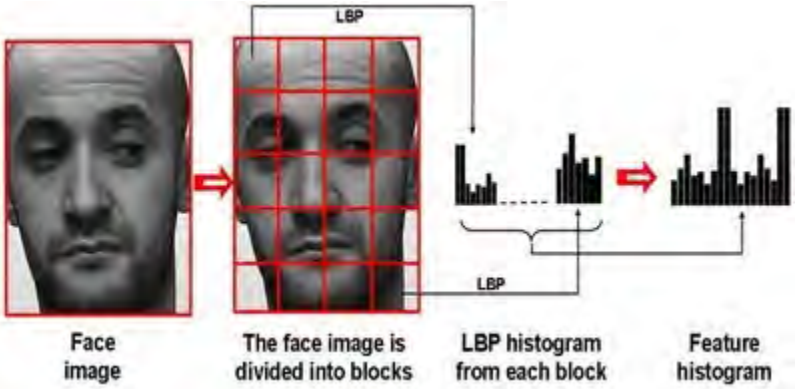
Textura

Local Binary Patterns (LBP)

Descriptores locales invariantes a cambios intensidad



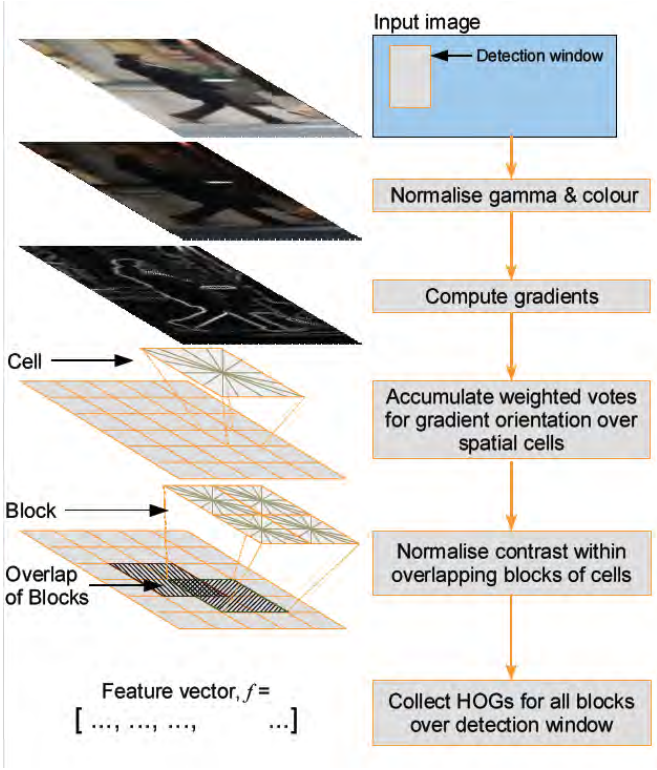


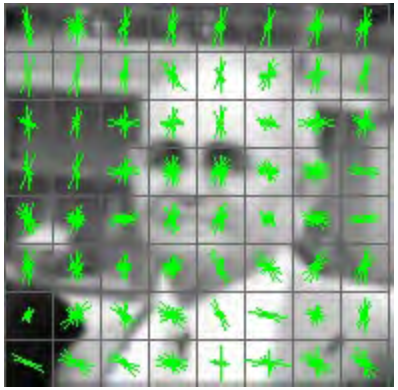


Caras

Textura

Histogram of Oriented Gradients (HOG)





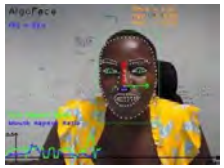
Caras

HOG + SVM

Elementos del rostro

Implementación en dlib

ALGOFACE



V. Kazemi and J. Sullivan. *One Millisecond Face Alignment with an Ensemble of Regression Trees*. CVPR, 2014

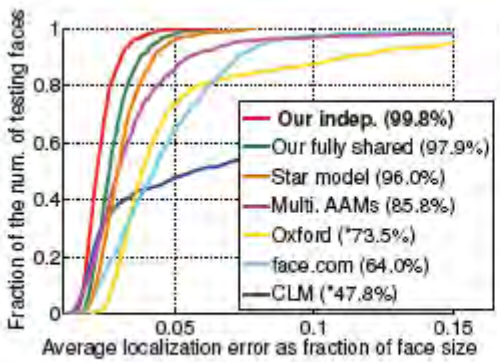
Caras

In the wild

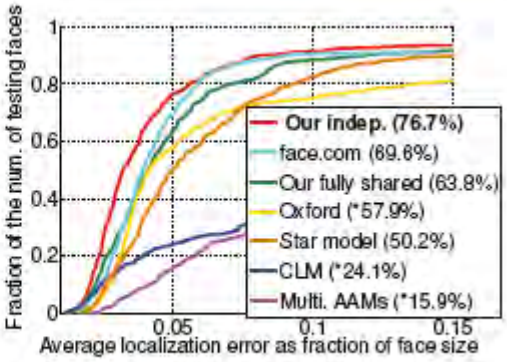
Combinaciones de árboles

Muestras anotadas manualmente con todos los puntos

Caras de cierto tamaño



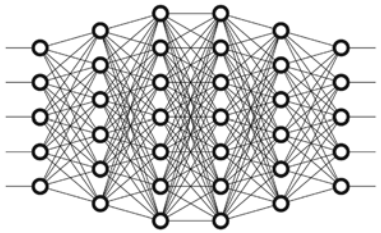
(a) MultiPIE



(b) AFW

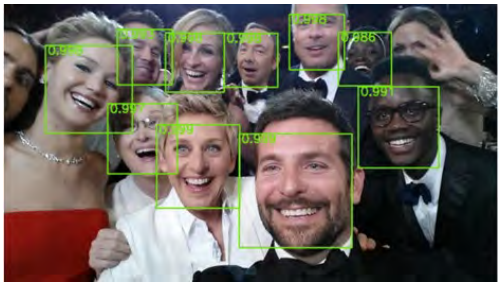
X. Zhu, D. Ramanan, "**Face Detection, Pose Estimation, and Landmark Localization in the Wild**", *Computer Vision and Pattern Recognition*, 2012

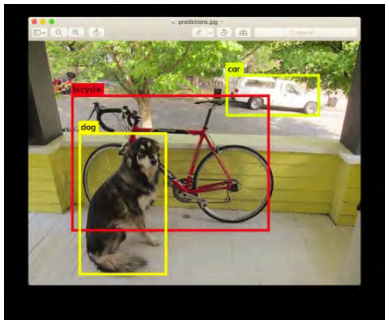
Caras



Convolutional Neural Networks (CNNs)

GPUs



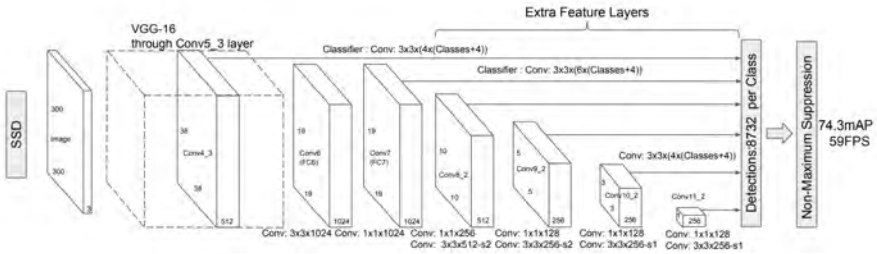


Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. *CVPR* 2013
Zhang, K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016

Caras

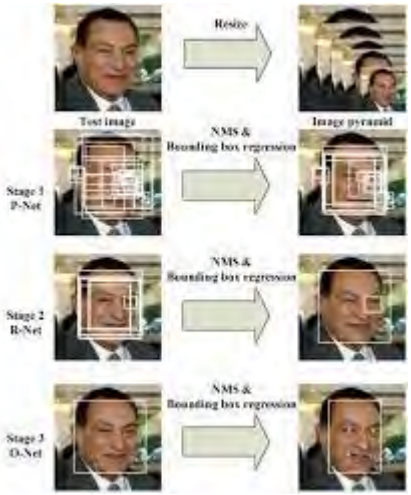
DL con código disponible:

OpenCV basado en SSD [1]



Dlib incluye un detector MMOD [2]

MTCNN [3]

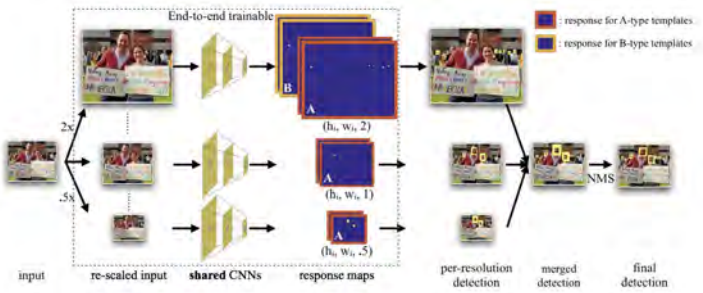


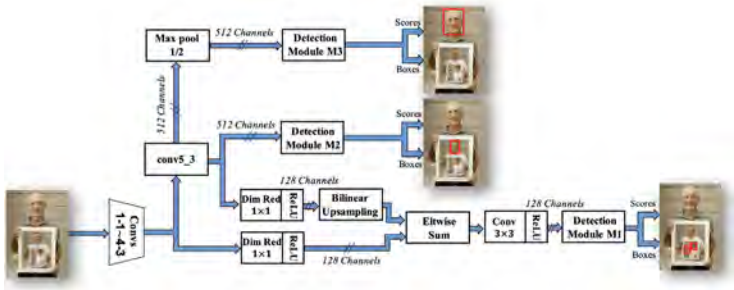
[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg. SSD: Single Shot MultiBox Detector. ECCV, 2016
[2] Davis E. King. Max-Margin Object Detection, arXiv:1502.00046, 2015
[3] Zhang, K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016

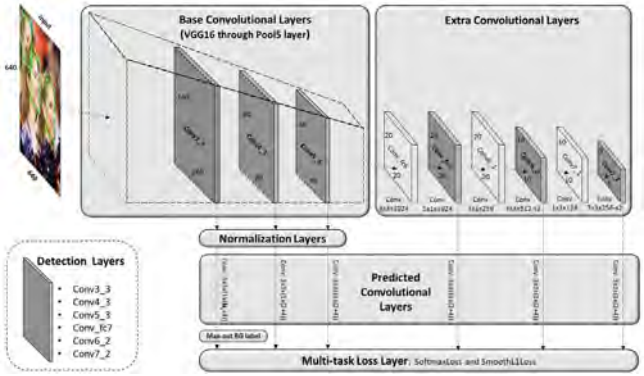
Caras

DL con código disponible:

Tiny face detector [4]
SSH [5]
FaceKit [6]
S3FD [7]
RetinaFace [8]
Img2face [9]







[4] P. Hu and D. Ramanan. Finding tiny faces. CVPR, 2017
[5] M. Najibi, P. Samangouei, R. Chellappa, Larry S. Davis SSH: Single Stage Headless Face Detector ICCV, 2017
[6] X. Shi, S. Shan, M. Kan, S. Wu, X. Chen, Real-Time Rotation-Invariant Face Detection with Progressive Calibration Networks. 2018
[7] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan Z. Li. S3FD: Single Shot Scale-invariant Face Detector ICCV, 2017
[8] J. Deng, J. Guo, Y. Zhou, J. Yu, Irene Kotsia, Stefanos Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild. arXiv:1905.00641, 2019
[9] . Albiero, X. Chen, X. Yin, G. Pang, T. Hassner, *img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation*, 2020

Caras

Retinaface

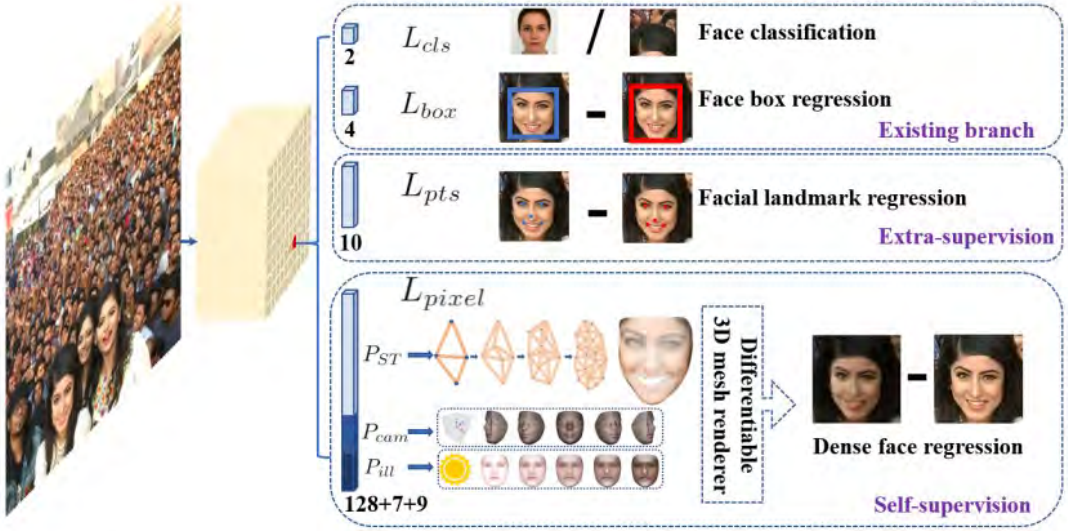
Pirámide: características (ResNet) a cinco escalas

Contexto de la cabeza

Regresión para estimar el contenedor

Multi tarea:

- Detección
- Máscara 2D
- Pose 3D



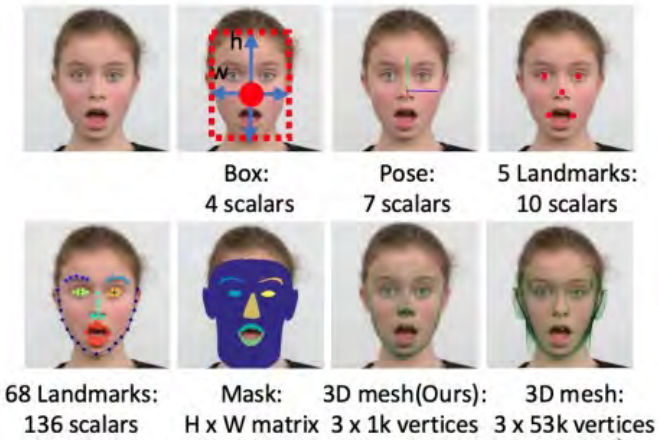
J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. 2019. *RetinaFace: Single-stageDense Face Localisation in the Wild.* 2019

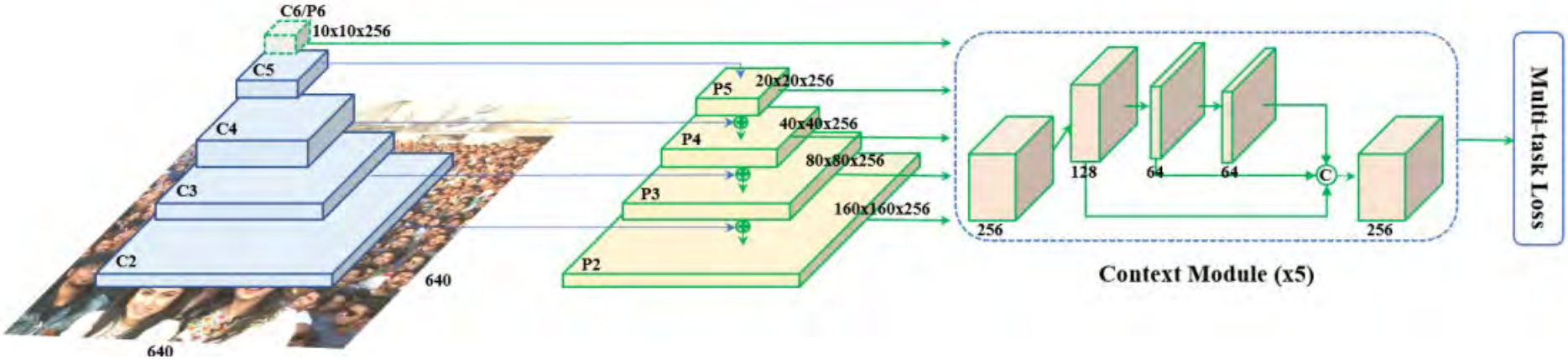
Caras

Retinaface

Multi tarea:

- Detección
- Máscara 2D
- Pose 3D





Caras

Benchmarks

[Face Detection Data Set (FDDB)](#)

[WiderFace](#)

[Unconstrained Face Detection Dataset](#) (UFDD)

[VGGFace2](#)

Caras

Benchmarks

TABLE 1
The performance of face detection models on different versions of Wider-Face dataset.

Method	Easy	Medium	Hard
Faceness [64]	71.3	63.4	34.5
Multiscale Cascade CNN [6]	69.1	66.4	42.4
CMS-RCNN [34]	90.2	87.4	64.3
LFFD [68]	89.6	86.5	77.0
img2pose [84]	90.0	89.1	83.9
S3FD [40]	92.8	91.3	84.4
EXTD [62]	91.2	90.3	85.0
FACE R-FCN [35]	94.3	93.1	87.6
SRN [50]	95.9	94.8	89.6
FDNet [37]	95.0	93.9	89.6
DSFD [51]	96.0	95.3	90.0
PyramidBox [49]	95.6	94.6	90.0
AInnoFace [69]	96.5	95.7	91.2
RetinaFace [53]	-	-	91.4
TinaFace [54]	-	-	92.4

TABLE 2
The performance of face detection models on the FDDB dataset.

Method	AP
CascadeCNN [25]	85.7
Joint-Cascade [28]	86.3
HyperFace [63]	90.1
Faceness [64]	90.3
DP2MFD [50]	90.3
UnitBox [70]	95.1
FaceBoxes [41]	96.0
Faster R-CNN [17]	96.1
DPSSD [85]	96.1
LFFD [68]	97.3
S3FD [40]	98.3
PyramidBox [49]	98.7
SRN [50]	98.8
FACE R-FCN [35]	99.0
DSFD [51]	99.1

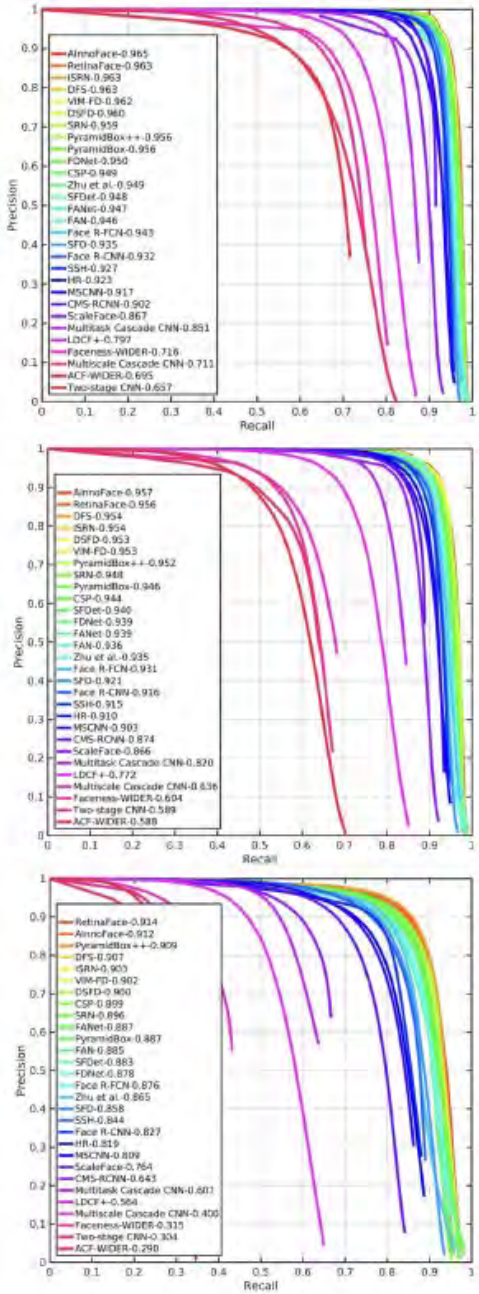


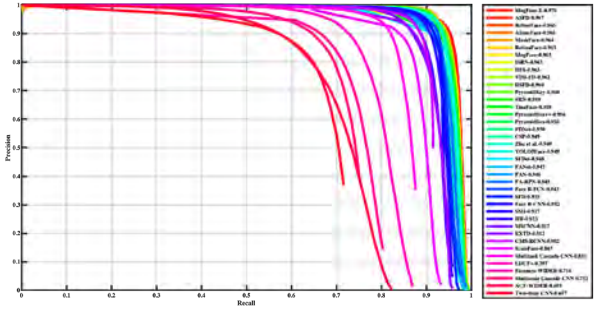
Fig. 35. The top, middle, and bottom figures present the Precision-Recall curves of promising face-detection models on WiderFace Easy/Medium/Hard test sets, respectively. Courtesy of [43].

Going Deeper Into Face Detection: A Survey, 2021

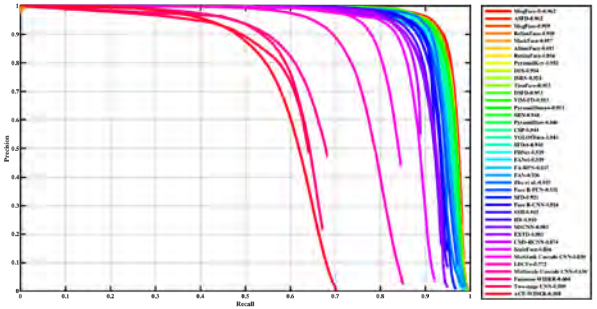
Caras

Benchmarks

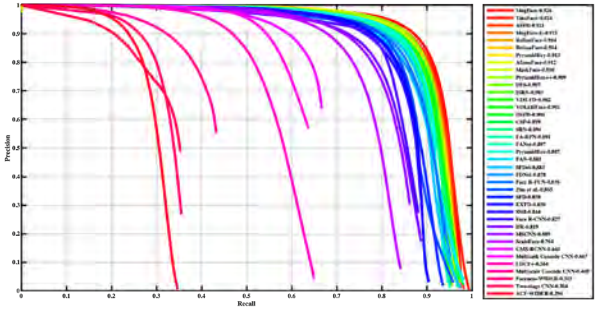
Wider Face (10/24)



Easy



Medium



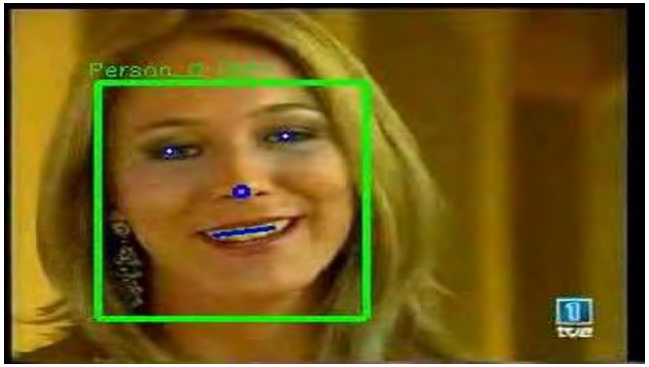
Hard

Caras

Comparativas de detectores

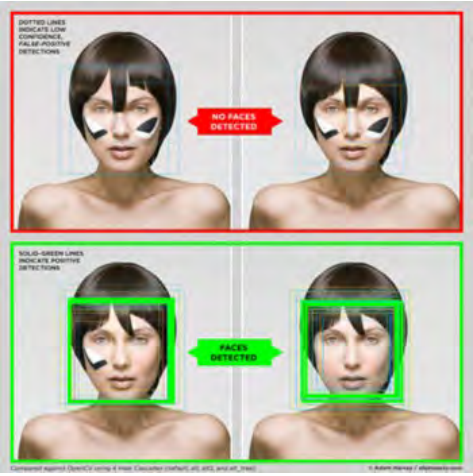
Detección en vídeo

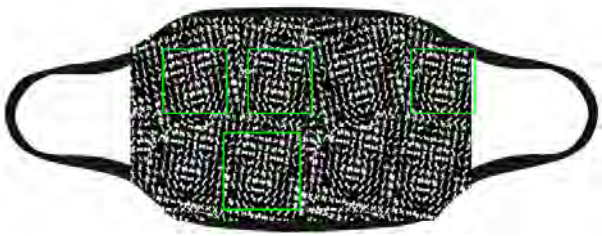
Tracking by detection



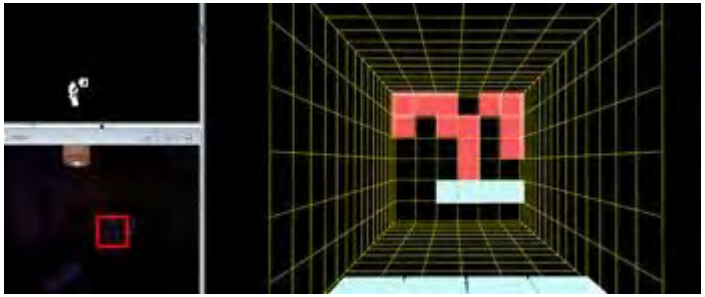
Caras

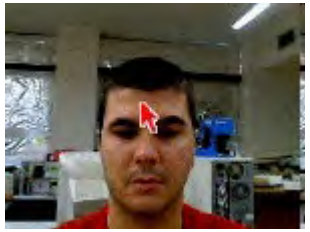
Confundir al detector
Privacidad



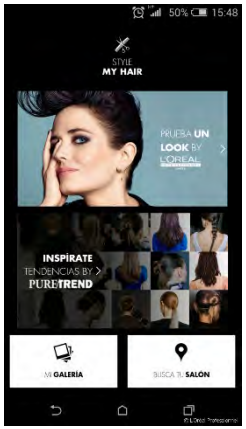


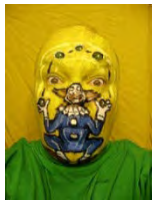
Aplicaciones





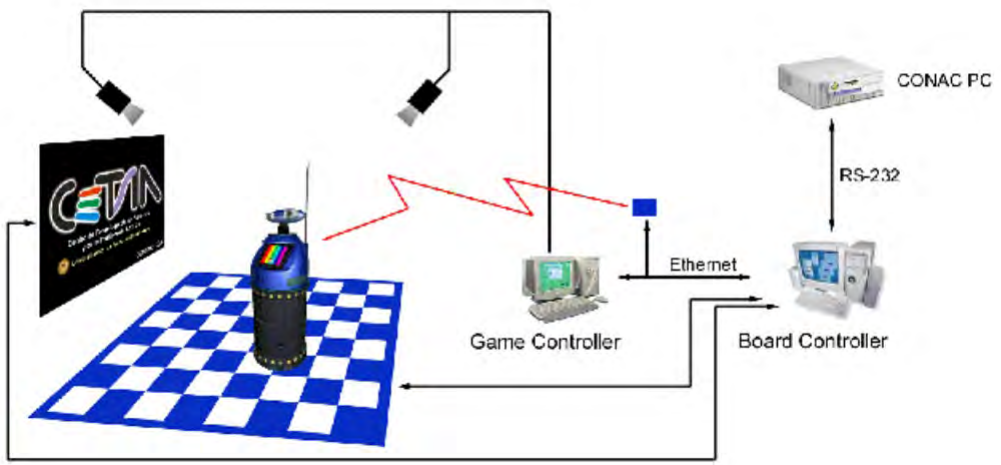
NISSAN





Personas

[Batallita] eldi



A. Domínguez et al., *Eldi: an agent based museum robot*, Systems Science 2001

Personas

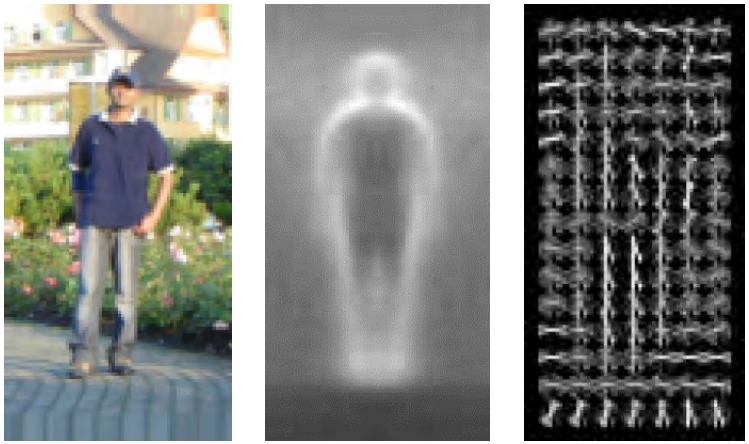
Dalal y Triggs (*pedestrian*) >39k citas
Histograma de gradientes orientados (HOG)
Máquinas de vectores soporte (SVM)

Dalal & Triggs, Histograms of Oriented Gradients, CVPR 2005

Personas

Dalal y Triggs, proceso:

1. Normalización color
2. Ventana deslizante (posiciones y escalas)
3. Modelo de tamaño fijo (64x128)
4. Gradientes, HOG en cada ventana 8x8 (9 bins)
5. Normaliza con vecinos
6. SVM lineal puntúa cada ventana
7. NMS para eliminar detecciones solapadas pobres

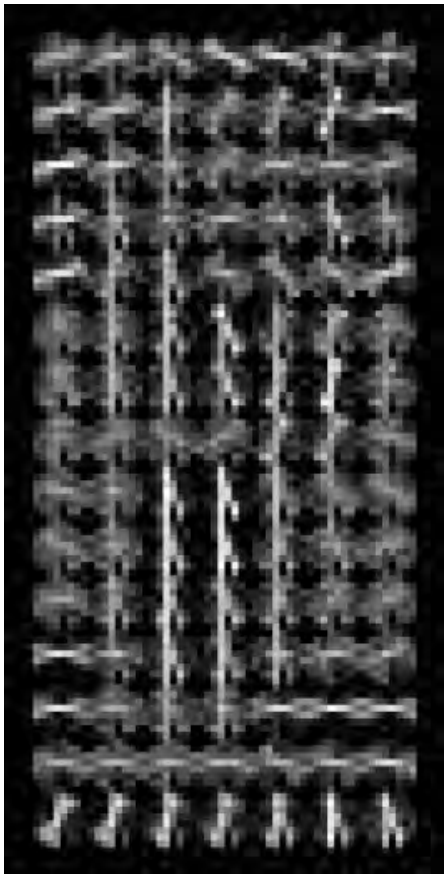


Adaptado de Pete Barnum

Personas

Dalal y Triggs

Características para ventana 8x8 en imagen 64x128



orientaciones

características = (15 * 7) * 9 * 4 = 3780

celdas

normalizaciones

Personas

Dalal y Triggs

Características

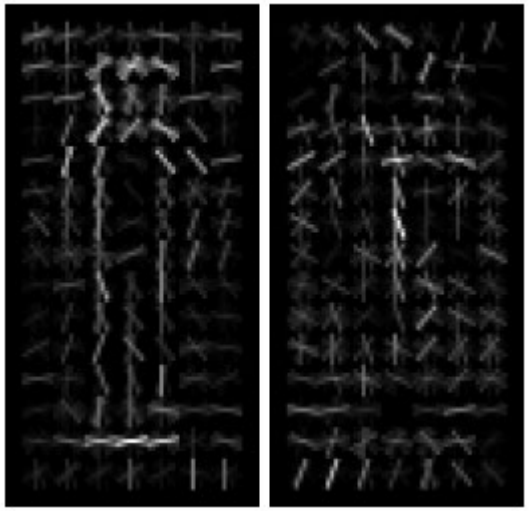
orientaciones

características = (15 * 7) * 9 * 4 = 3780

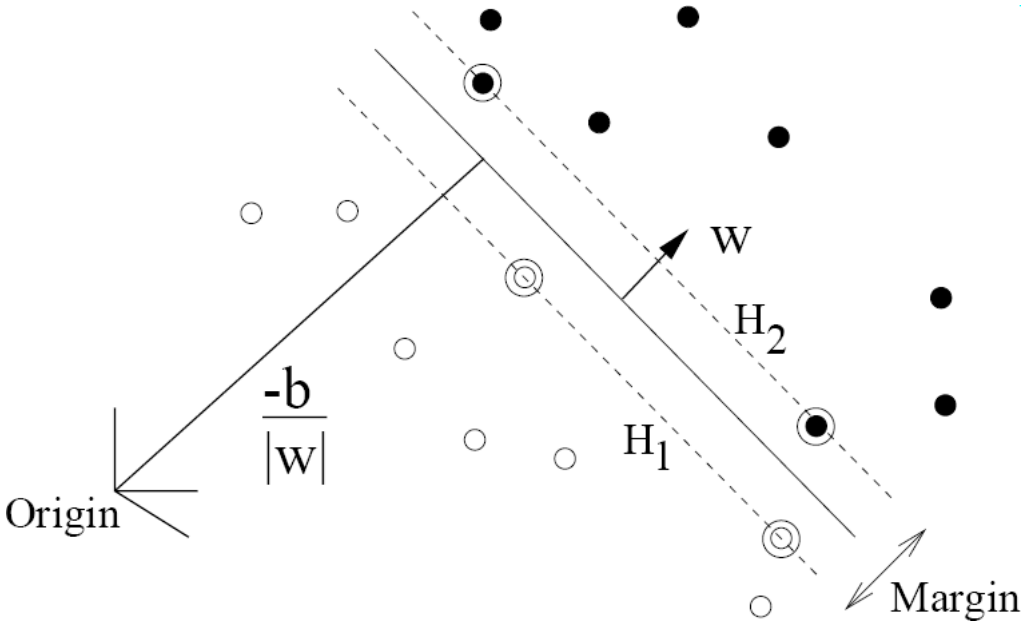
celdas

normalizaciones

Clasificador SVM

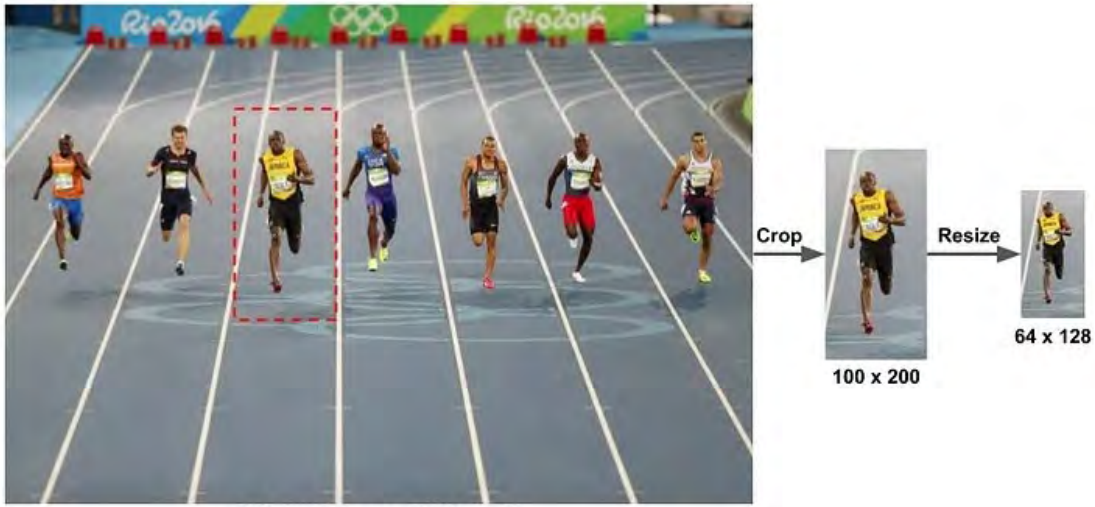


pos w neg w

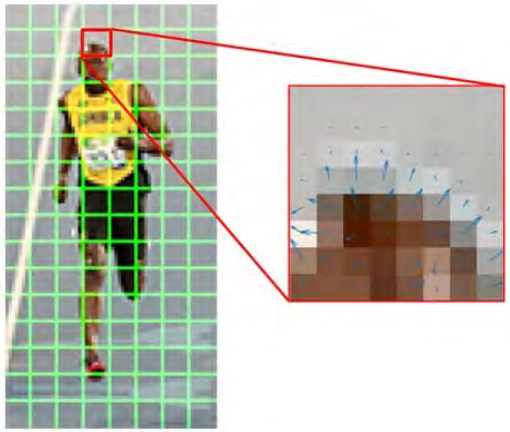


Personas

Dalal y Triggs



Original Image : 720 x 475



2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

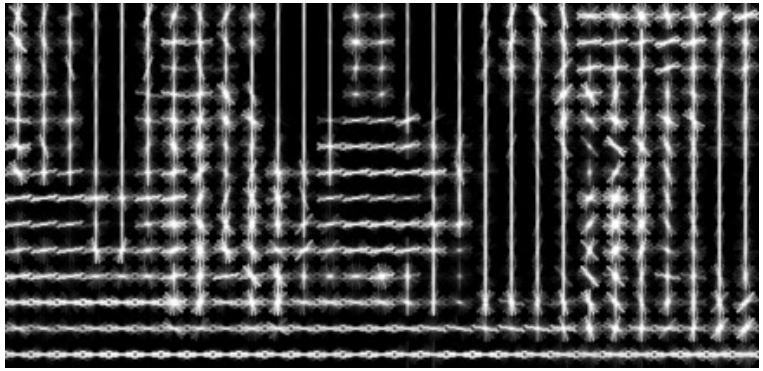
80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

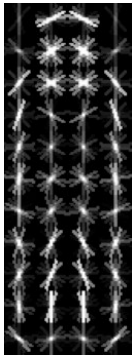
Personas

Dalal y Triggs

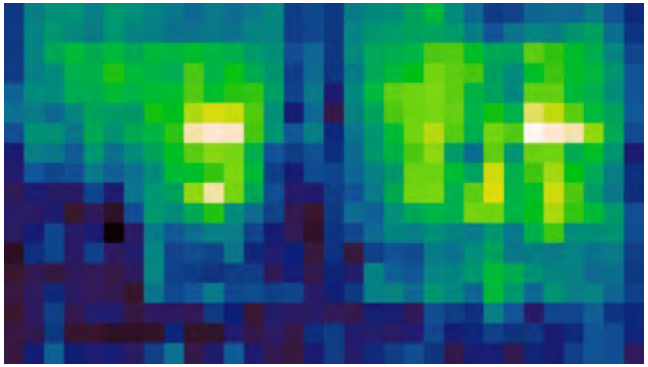
Mapa HOG



Modelo



Mapa de respuesta

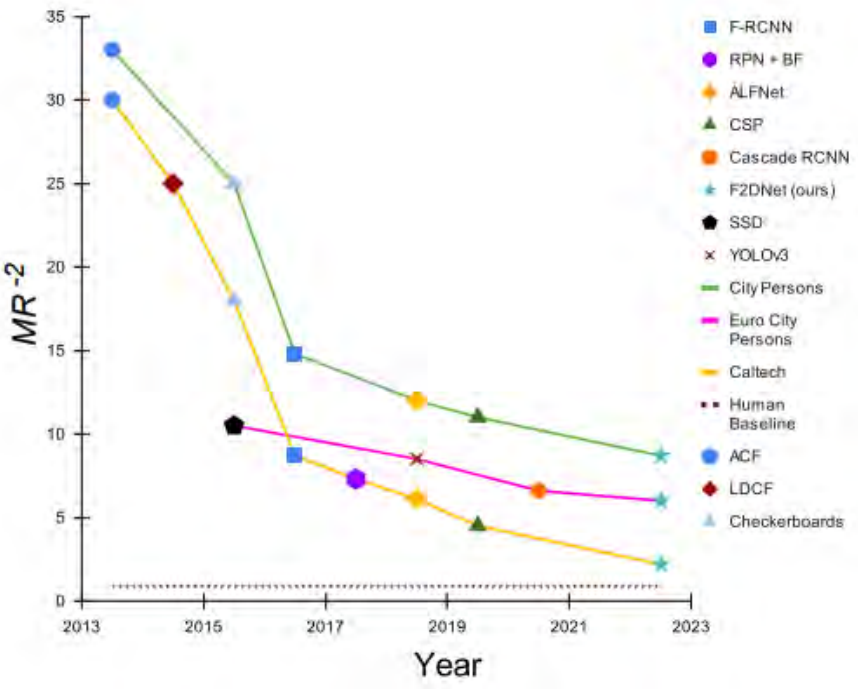


Pete Barnum

Personas

Benchmarks

- CityPersons
- EuroCityPersons
- Pedestrian Caltech



SOTA: FD2Net, EGCL

Varios sobre MMDetection (*open source object detection toolbox*)

DPM

Ventana deslizante y modelos estadísticos
Bien con caras, personas
Competitivo con coches, personas
Mal con gatos y perros

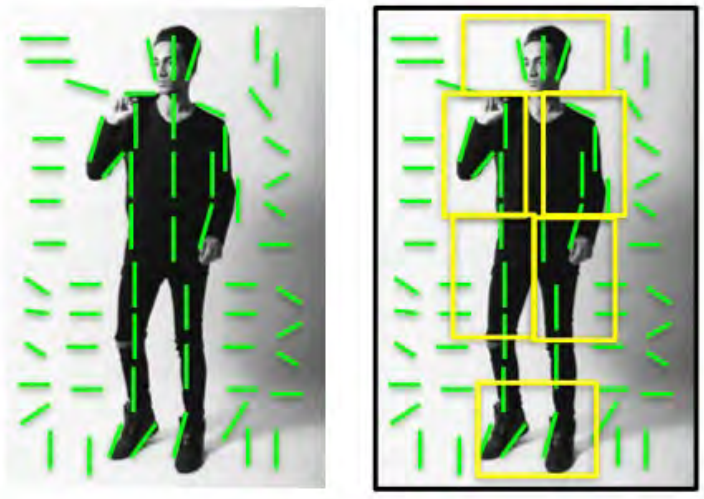
¿Por qué tan mal con algunas clases?

Bien con objetos no deformables
Vistas canónicas

Mejores clasificadores y características

DPM

Dalal y Triggs estructura rígida
Dificultad para asumir deformaciones de pose
Los humanos no son tan *rígidos*
Bag of words tolera deformaciones pero no detección fina
Otras categorías mucho menos
¿Cómo modelar mayores variaciones de apariencia?



DPM

Felzenszwalb et al. *Deformable part models* (DPM)

HOG, PCA para reducir dimensión

Latent SVM

Tolerar grandes variaciones

Mezcla de modelos

R. Girschick

P. F. Felzenszwalb et al. Object Detection with Discriminatively Trained Part-Based Models PAMI 2009

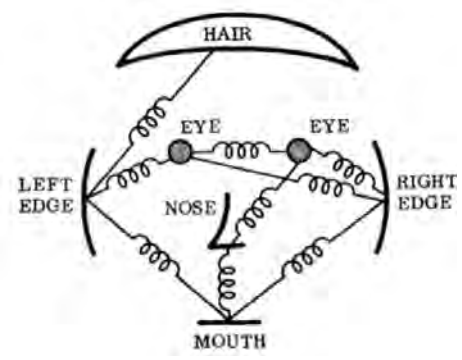
DPM

HOG, PCA para reducir dimensión

Latent SVM

Mezcla de modelos

- Modelo global
- Modelo de partes y conexiones



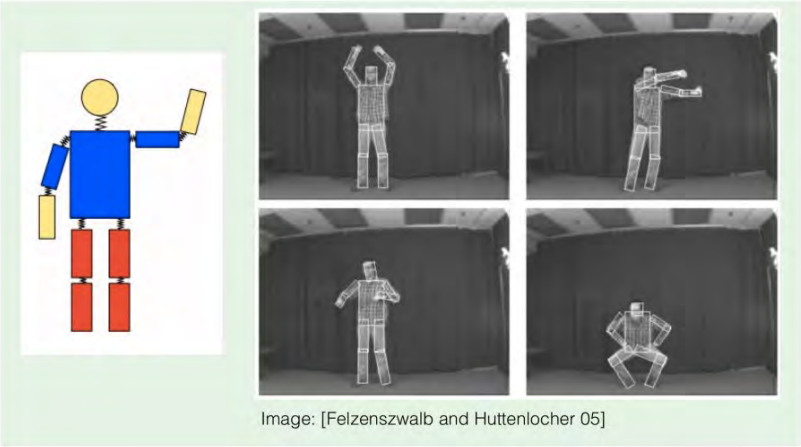
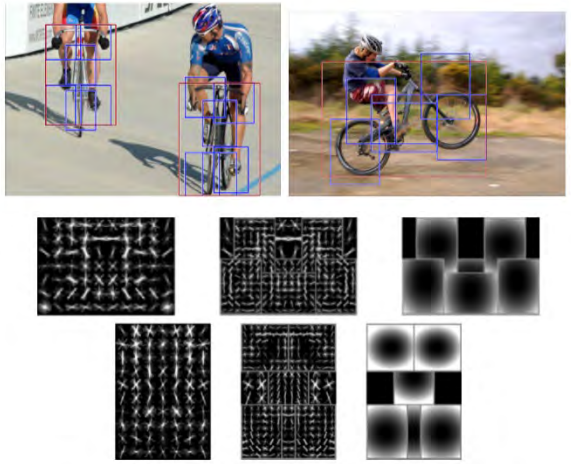


Image: [Felzenszwalb and Huttenlocher 05]



P. F. Felzenszwalb et al. Object Detection with Discriminatively Trained Part-Based Models PAMI 2009

DPM

Grafo G=(V,E)

Partes $V = \{v_1, \dots, v_n\}$

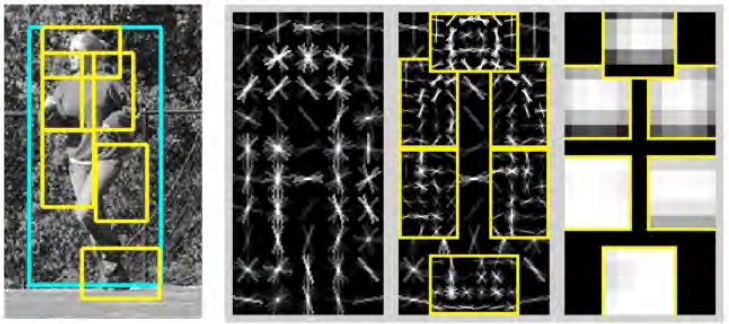
Si $(v_i, v_j) \in E$ hay conexión entre las partes

Configuración óptima es $L=(l_1,\dots,l_n)$ que maximiza

$$E(L) = \sum_{1}^{n} m_i(l_i) - \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j)$$

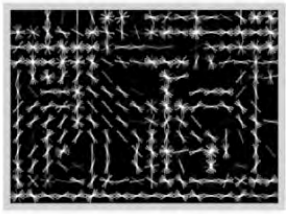
$m_i(l_i)$ coste de colocar la parte i en la posición l_i

$d_{ij}(l_i, l_j)$ coste de deformación



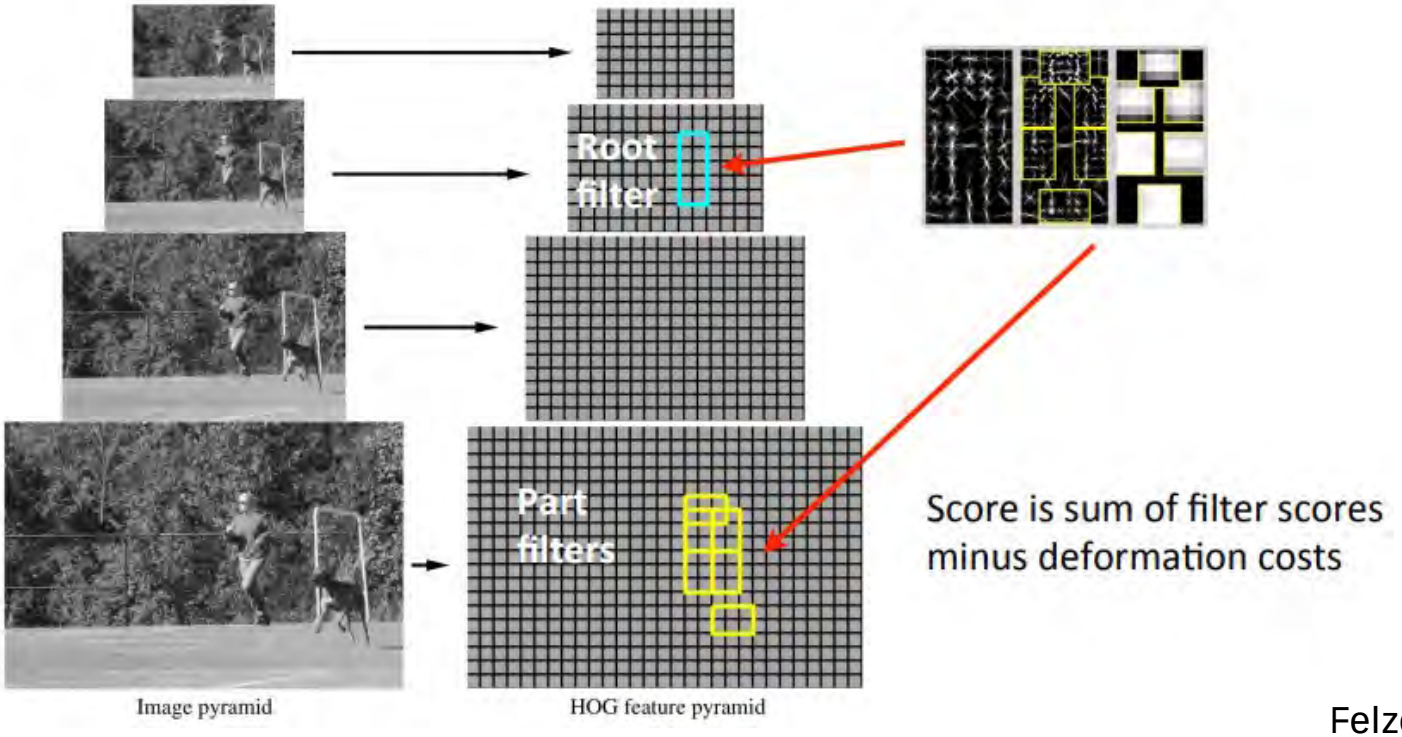
detection root filter part filters deformation models

DPM



Proceso:

- Imagen en bloques 8x8, HOG
- Diferentes resoluciones
- Filtro global y de partes (con deformación)
- Puntúa ventanas con filtros global, de partes y deformación

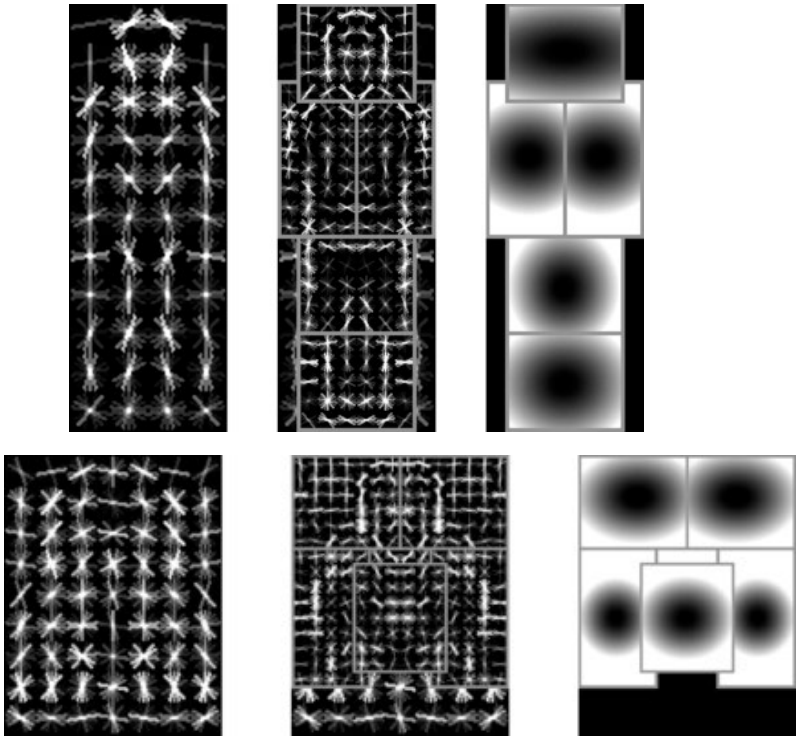


Felzenszwalb

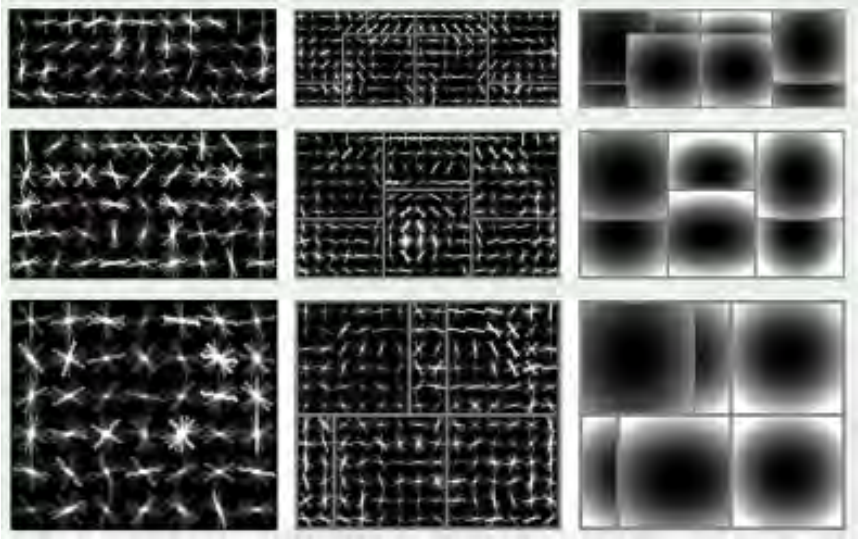
DPM

Mezcla de modelos

Persona (2)

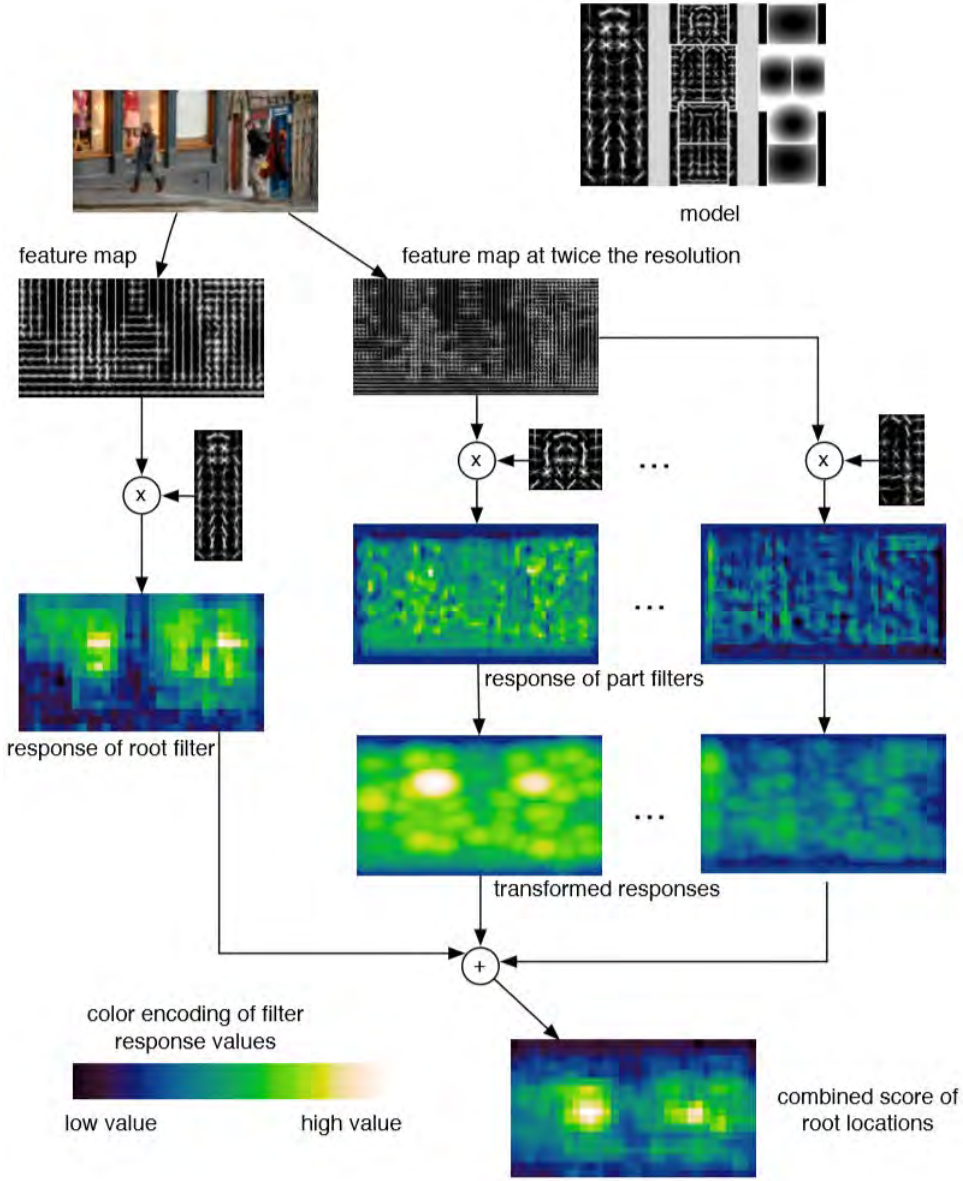


Coche (3)



Felzenszwalb

DPM



Felzenszwalb

Detectores actuales

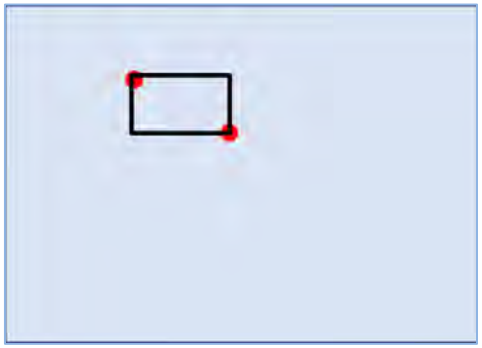
Detección como problema de clasificación:

- Clasificar cada posible ventana/caja

- ¿Número de cajas? Dos píxeles componen una

- Imagen 500x500, píxeles n=250000

$$\binom{n}{2} = \mathrm{o}(n^2)$$



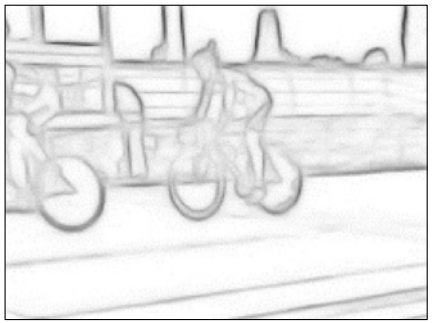
- Se dispara, mayoría de las cajas negativas

Detección como clasificación

Búsqueda selectiva:

1. Selección ventanas candidatas

Imagen

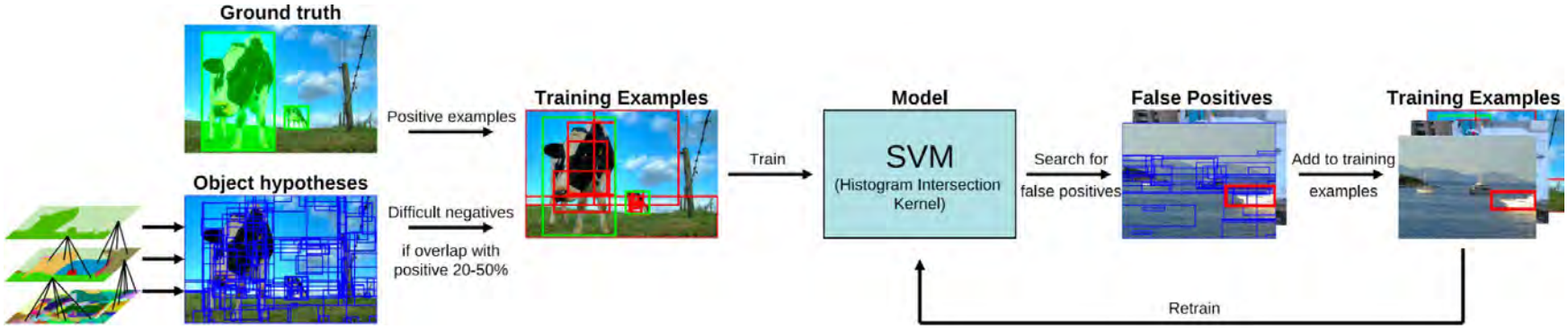


Bordes

Candidata

Zitnick y Dollar

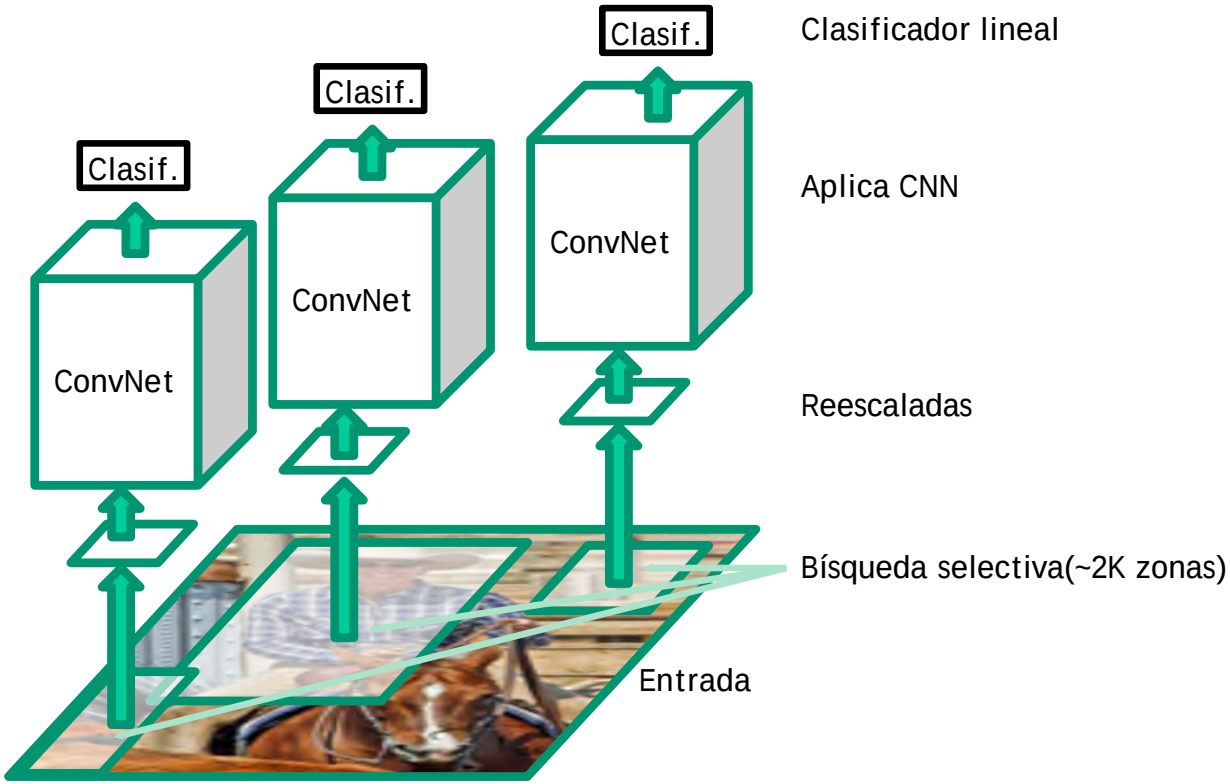
2. Clasificación sólo en candidatas



Uijlings et al.

R-CNN

Propuesta de regiones y características CNN



R. Girshick et al. **Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation**, CVPR 2014

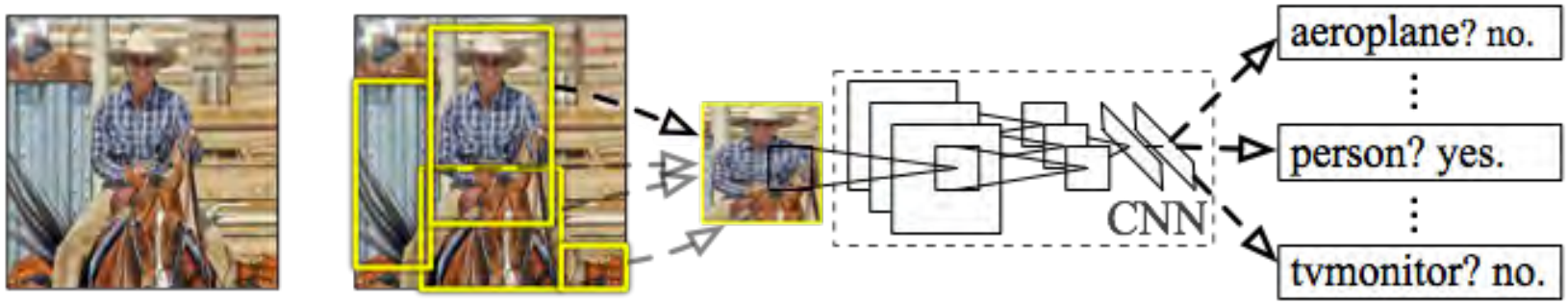
R-CNN

Propuesta de regiones y características CNN

- Escalado a tamaño fijo 227x227 (sin relación aspecto)
- CNN, características FC7, vector 4096 valores
- Clasificadores lineales (p.e. softmax, SVM, …)
- Regresión contenedor (revisa localización caja)
- NMS

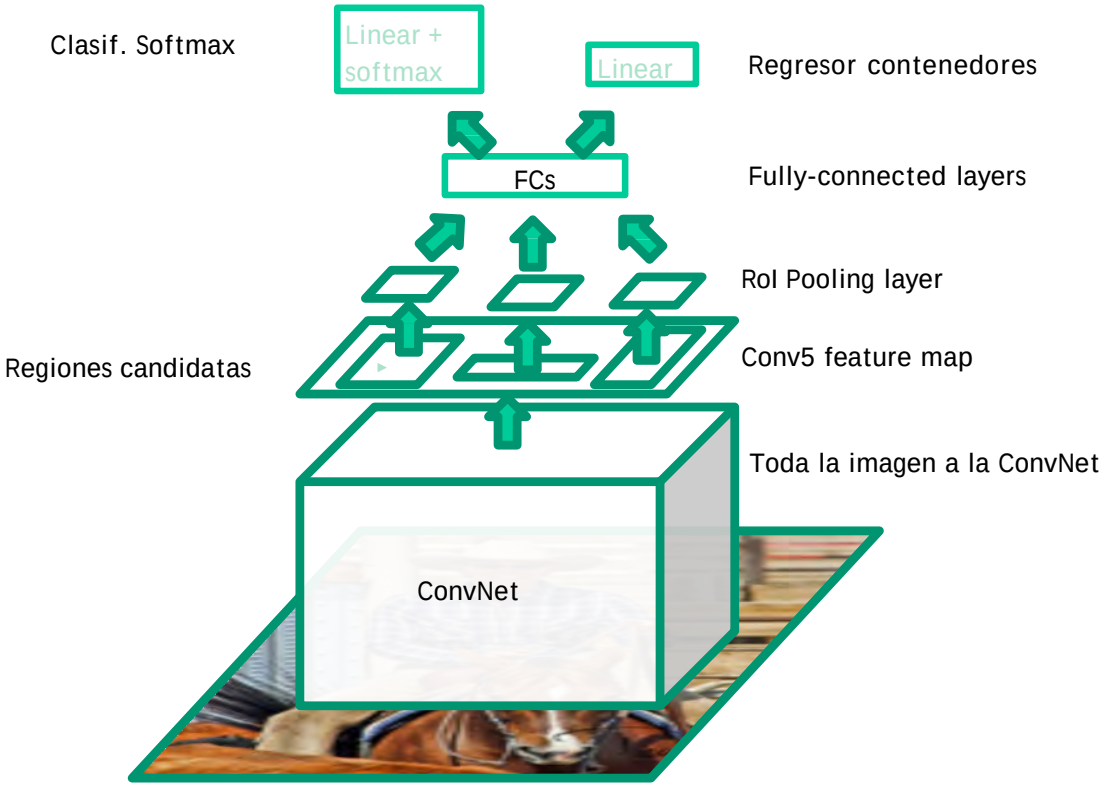
Lento

Selección de ventana no óptima



Fast R-CNN

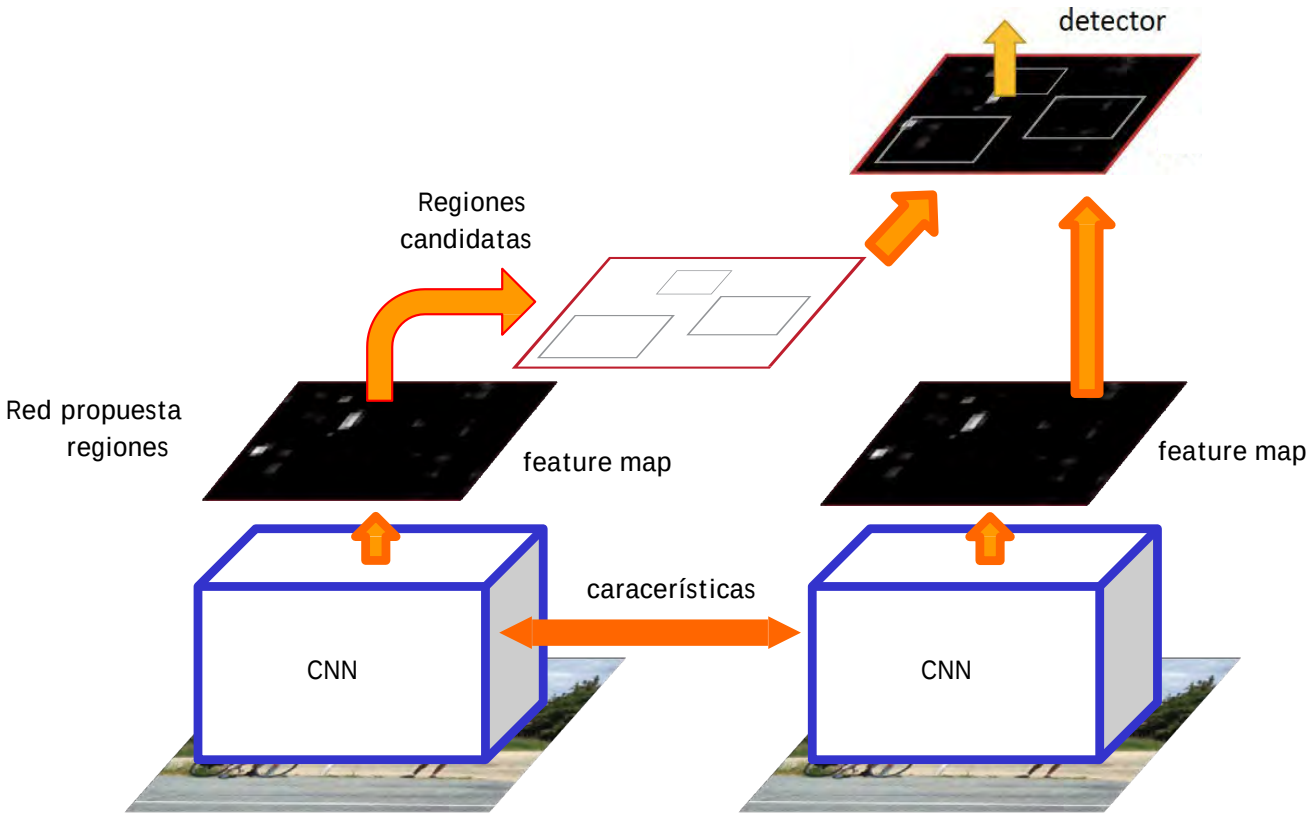
Optimiza
Reutiliza características entre candidatas
Extrae candidatas desde el mapa de características



R. Girshick et al., Fast R-CNN, ICCV 2015

Faster R-CNN

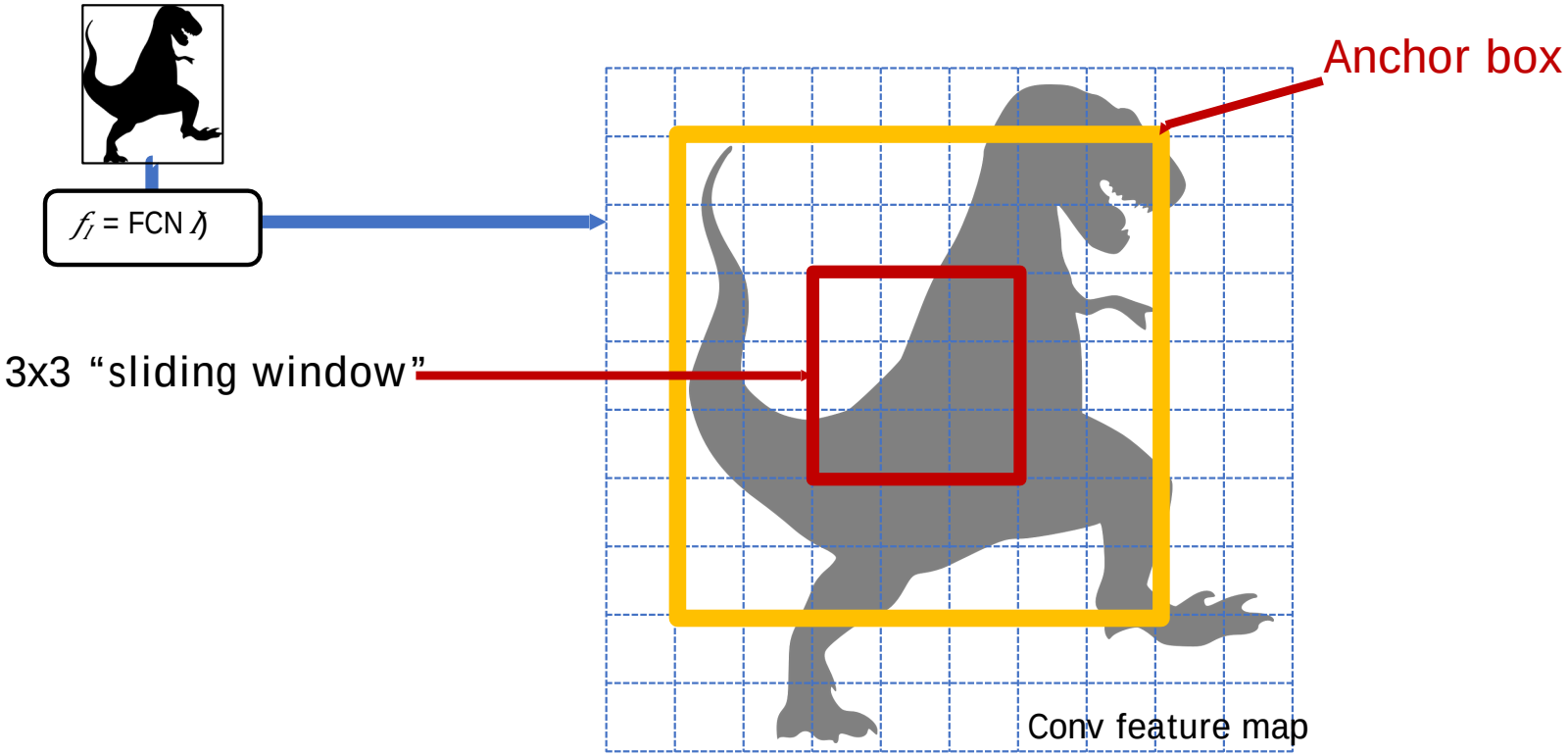
Entrena propuesta de regiones candidatas



S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015

Faster R-CNN

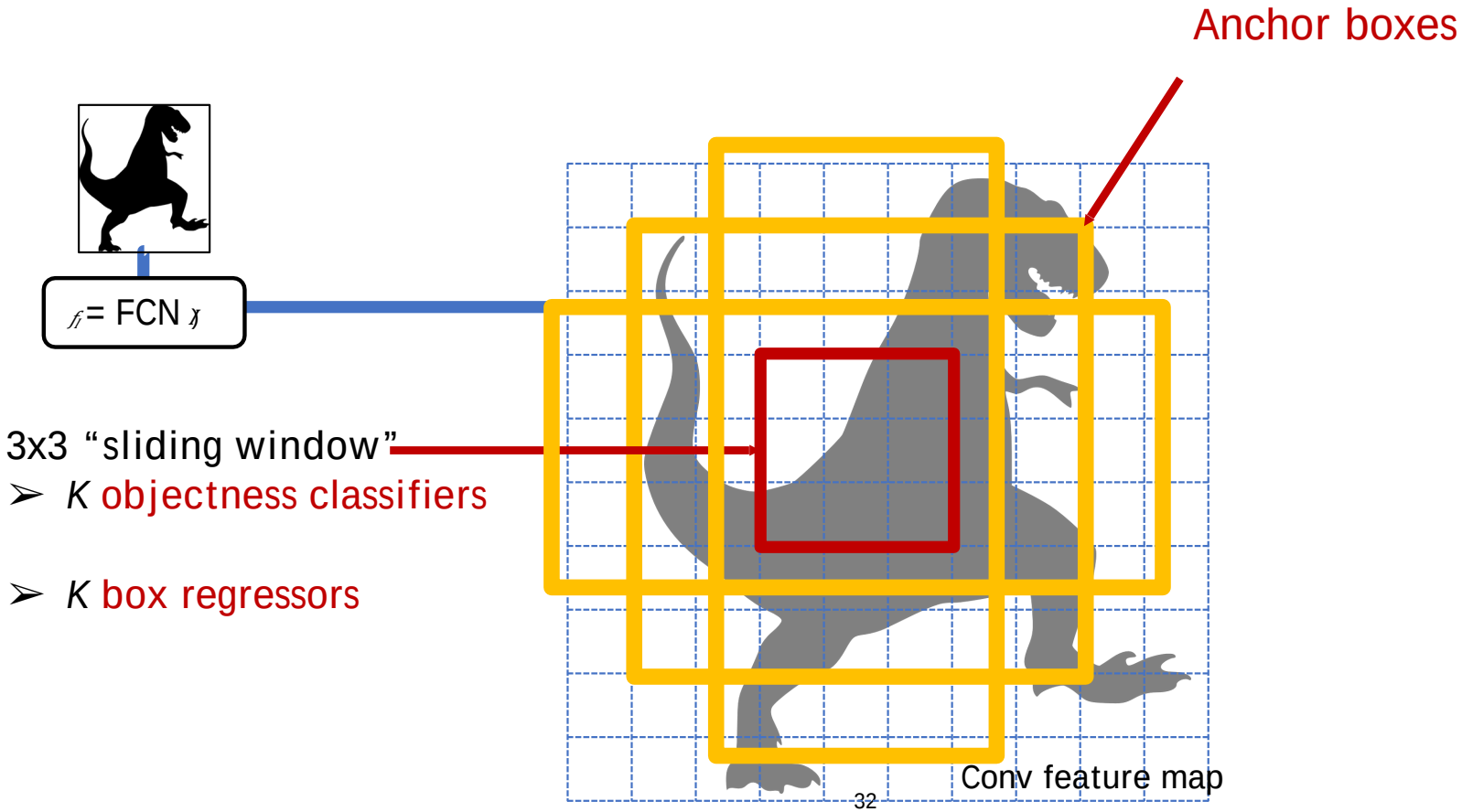
Red de regiones candidatas
Ventana deslizante 3x3 buscando objetos
Puntúa *objectness* (0,1)
Regresor tamaño (*anchor box*)



R. Girshick

Faster R-CNN

K *anchor boxes* por posición: diferentes tamaños y aspectos
K clasificadores de *objectness*



R. Girshick

Comparativa

K *anchor boxes* por posición: diferentes tamaños y aspectos
K clasificadores de *objectness*

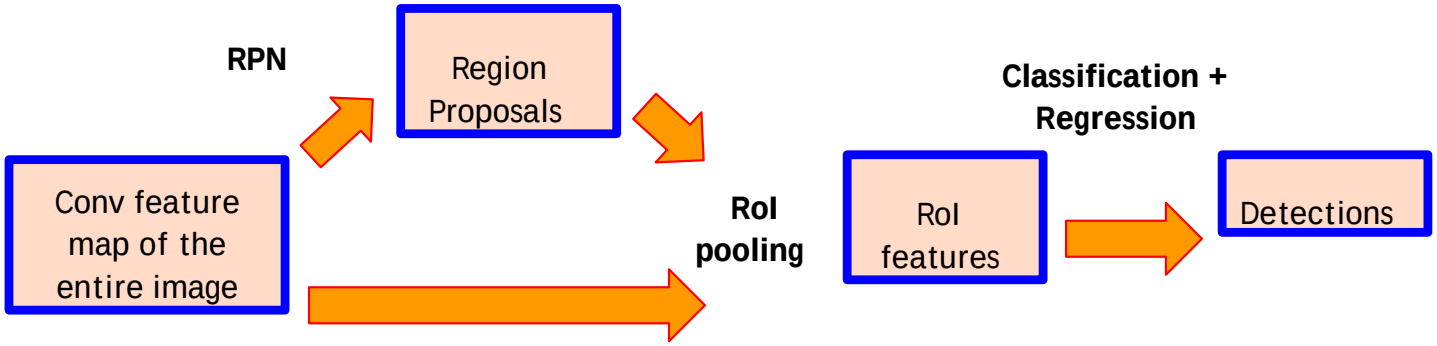
system	time	07 data	07+12 data
R-CNN	~50s	66.0	-
Fast R-CNN	~2s	66.9	70.0
Faster R-CNN	198ms	**69.9**	**73.2**

detection mAP on PASCAL VOC 2007, with VGG-16 pre-trained on ImageNet

Source: S. Lazebnik

De una etapa

Faster R-CNN separa propuesta de regiones de clasificación



¿Concentrarlo en una etapa?

S. Lazebnik

Ejemplos:

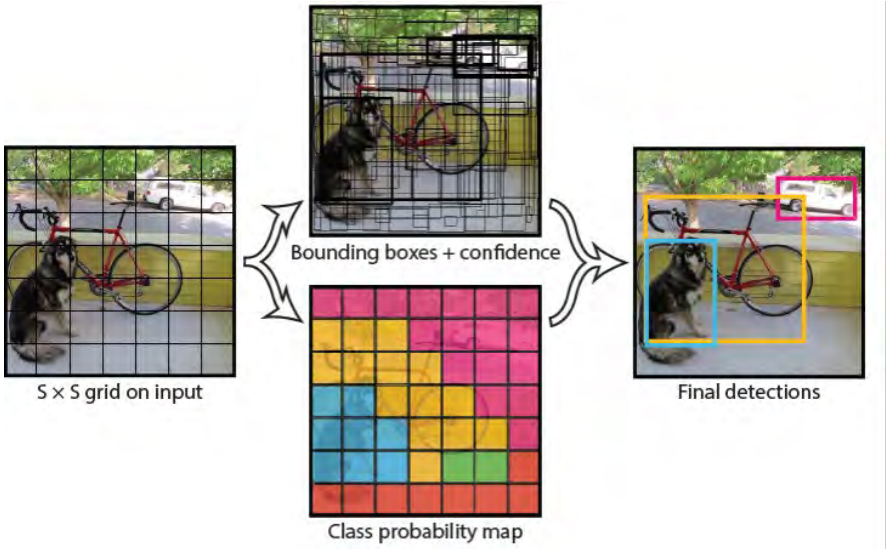
- SSD (*Single Shot MultiBox Detector*)
- RetinaNet
- YOLO (*You Only Look Once*)

YOLO

Divide imagen (448x448) en rejilla *gruesa* de SxS celdas
Detecta objeto en celda con centro (IoU)
Para cada celda:

- Puntuación de cada clase, predice etiqueta
- Contenedores candidatos, máximo dos por celda

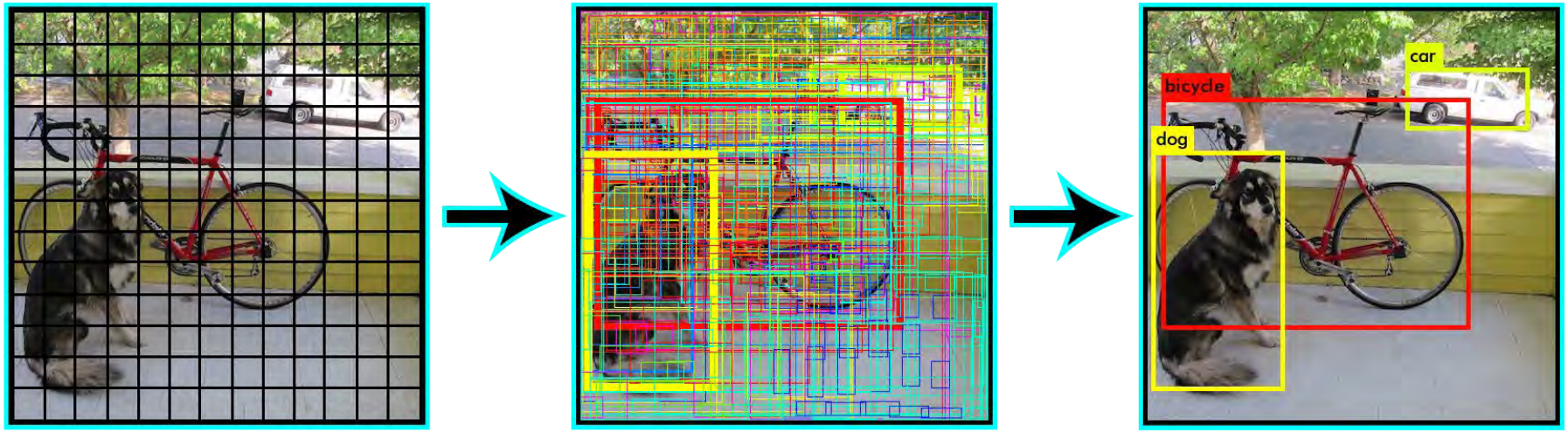
Objeto detectado múltiples veces, NMS por confianza



J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016

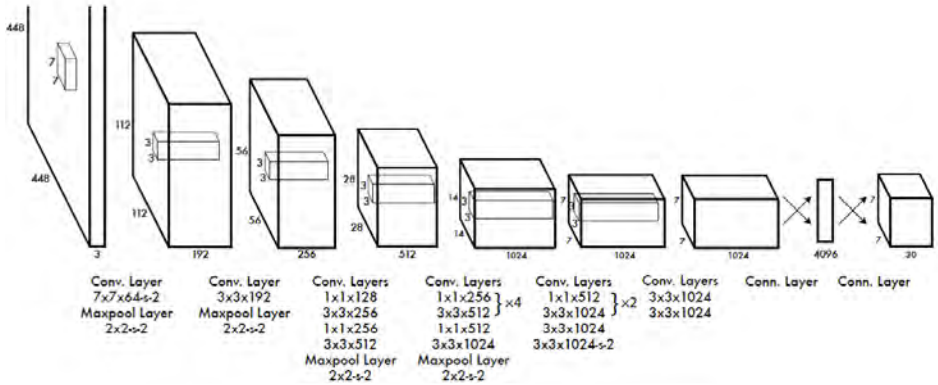
YOLO

Mismo objeto detectado en múltiples celdas, NMS
Un único objeto por celda
Objetos pequeños detectables limitado por tamaño celda



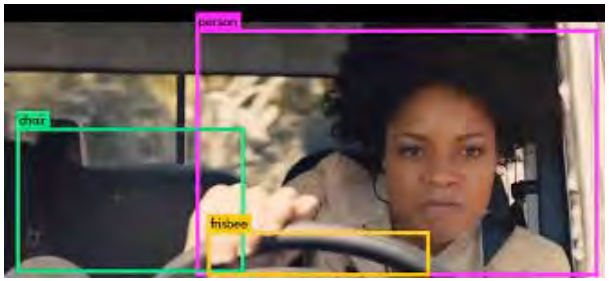
YOLO

Original, inspirada en GoogleNet 24 convolucionales, 2 FC
Mapa de características entrada 7x7



Pascal VOC

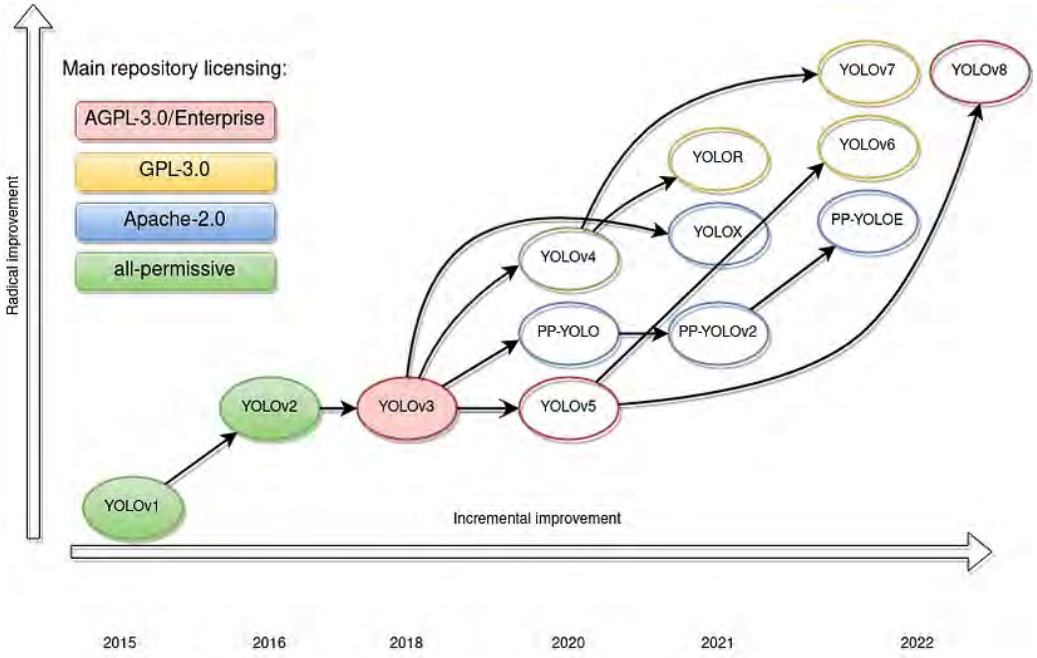
- 7x más rápido que Faster R-CNN (45-155 fps vs. 7-18 fps)
- Menos exacto (65% vs. 72 mAP%)



traffic light
traffic light
traffic light
stop sign
car
car
car

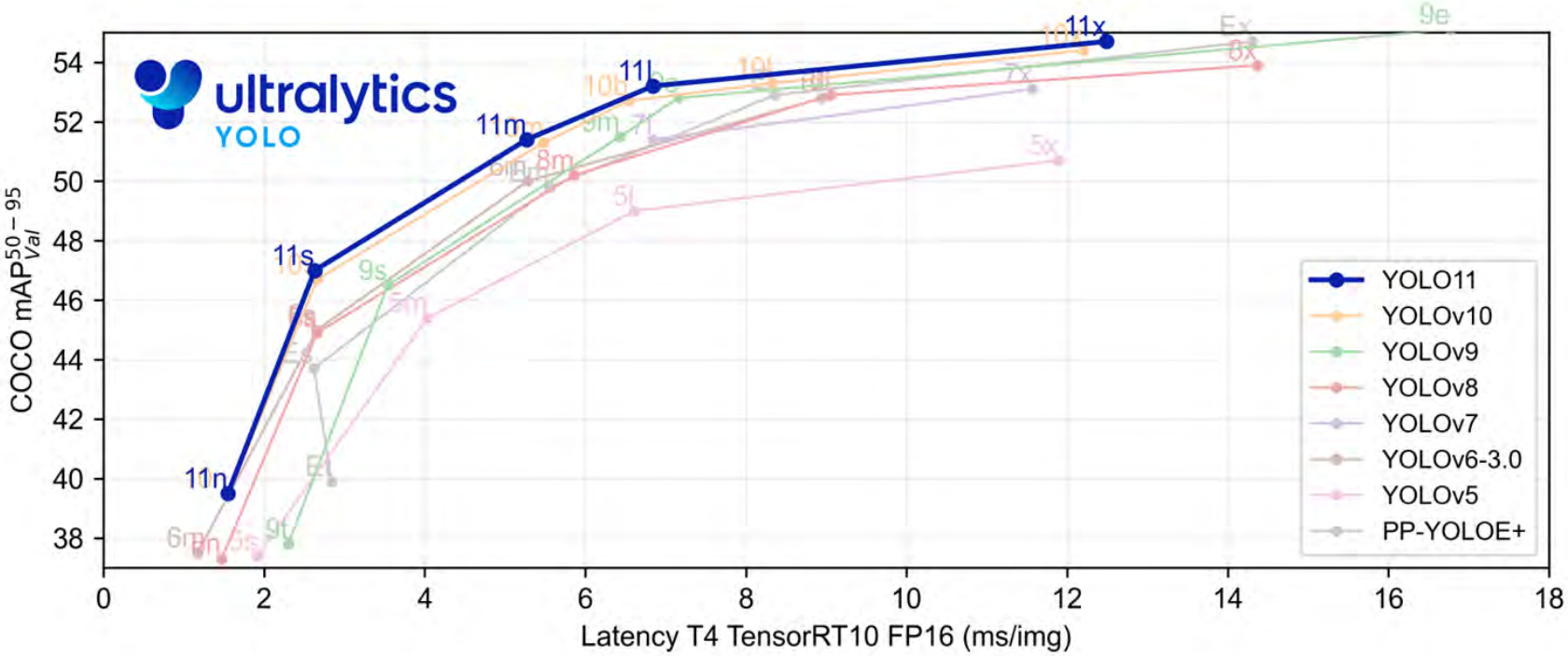
YOLO

Familia YOLO, cambiando elementos como arquitectura



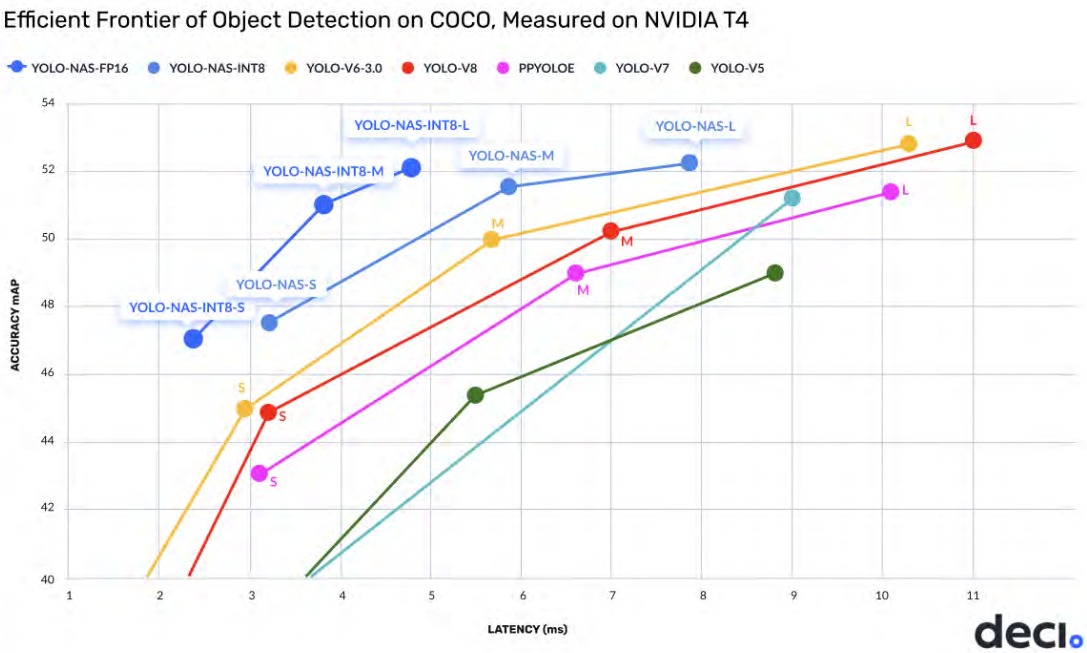
YOLO

Familia YOLO
9/24



YOLO

Familia
YOLO-NAS destaca



Model	mAP	Latency (ms)
YOLO-NAS S	47.5	3.21
YOLO-NAS M	51.55	5.85
YOLO-NAS L	52.22	7.87
YOLO-NAS S INT-8	47.03	2.36
YOLO-NAS M INT-8	51.0	3.78
YOLO-NAS L INT-8	52.1	4.78
YOLOv8 S	47.0	4.0
YOLOv8 M	50.0	6.5
YOLOv8 L	52.5	9.5

YOLO

Joseph Redmon · Feb 20, 2020
@pjreddie · **Follow**
"We shouldn't have to think about the societal impact of our work because it's hard and other people can do it for us" is a really bad argument.

> **Roger Grosse** @RogerGrosse
> Replying to @kevin_zakka and @hardmaru
> To be clear, I don't think this is a positive step. Societal impacts of AI is a tough field, and there are researchers and organizations that study it professionally. Most authors do not have expertise in the area and won't do good enough scholarship to say something meaningful.

Joseph Redmon
@pjreddie · **Follow**

I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.

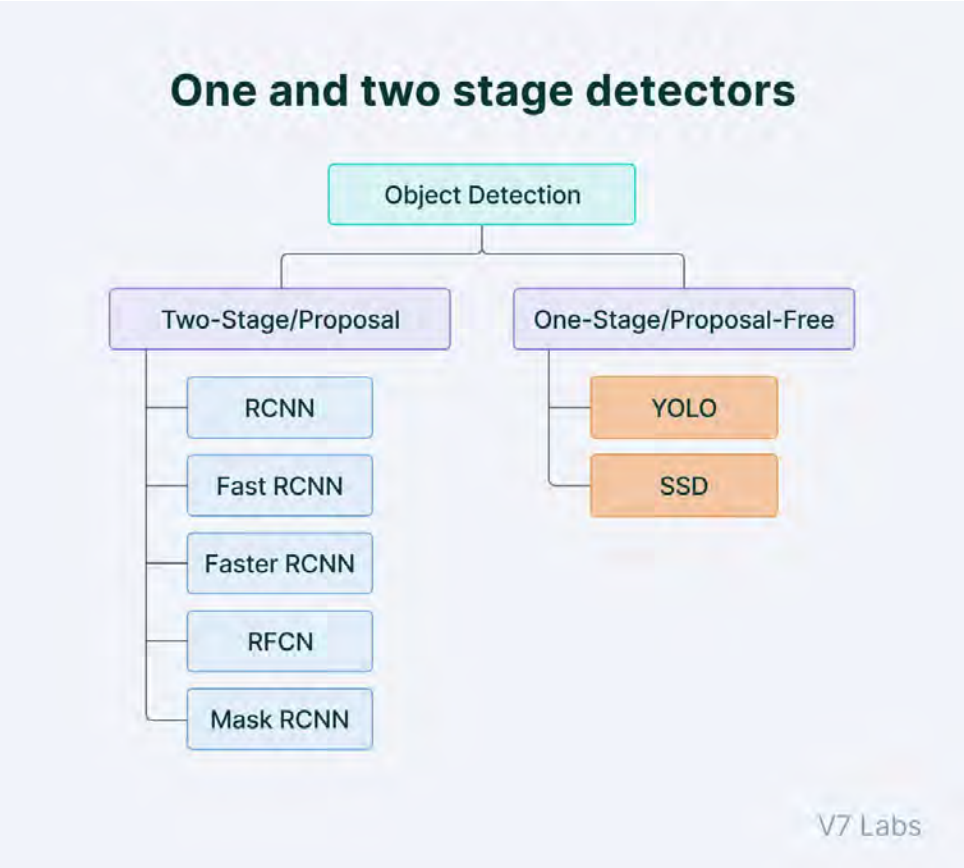
> **Roger Grosse** @RogerGrosse
> Replying to @skoularidou
> What's an example of a situation where you think someone should decide not to submit their paper due to Broader Impacts reasons?

5:09 PM · Feb 20, 2020

3.4K **Reply** **Share**

Una o dos etapas

Dos pasadas, primera para determinar localizaciones posibles
Una pasada, menor coste



YOLO

Detectores convolucionales

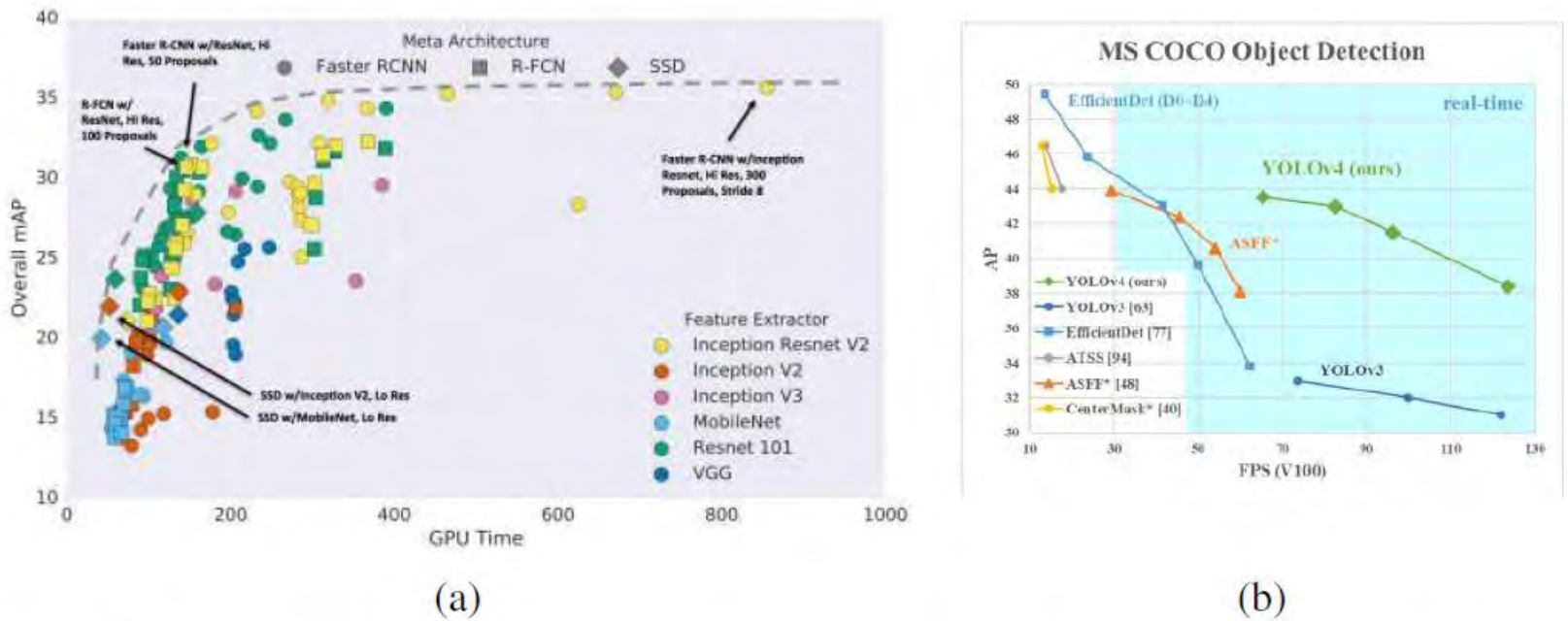


Figure 6.31 *Speed/accuracy trade-offs for convolutional object detectors: (a) (Huang, Rathod* et al. *2017) © 2017 IEEE; (b) YOLOv4 © Bochkovskiy, Wang, and Liao (2020).*

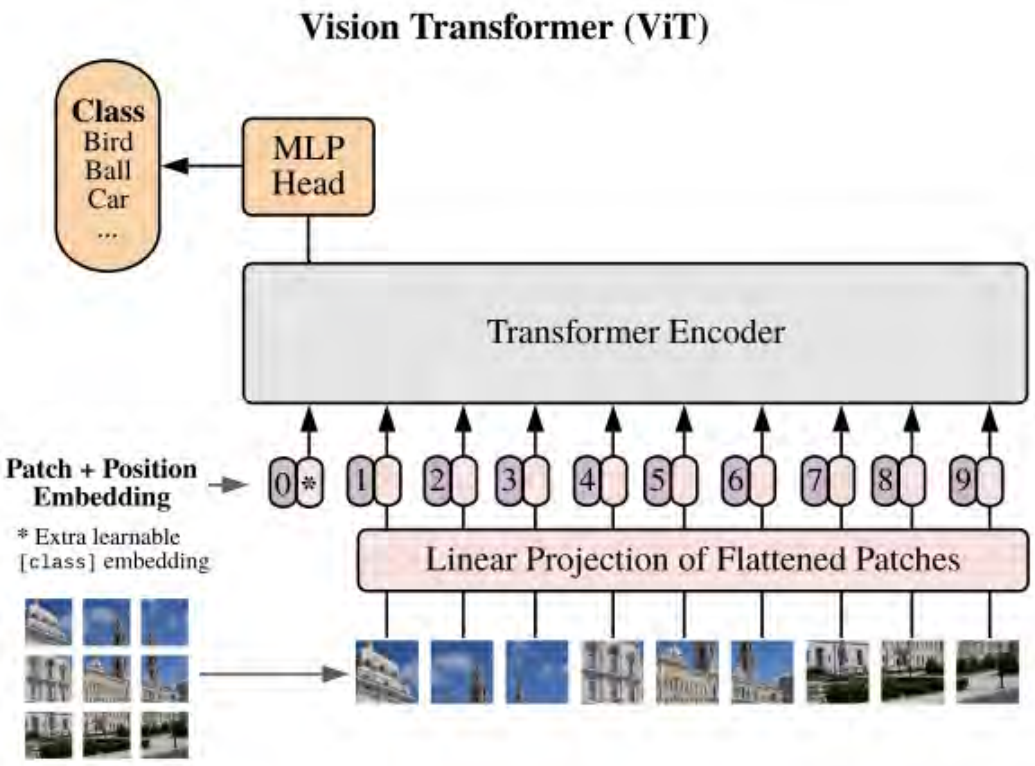
Transformers

Inicialmente NLP, desde 2017

Sustituir CNN y RNN

Convolucionales pierden contexto zonas lejanas

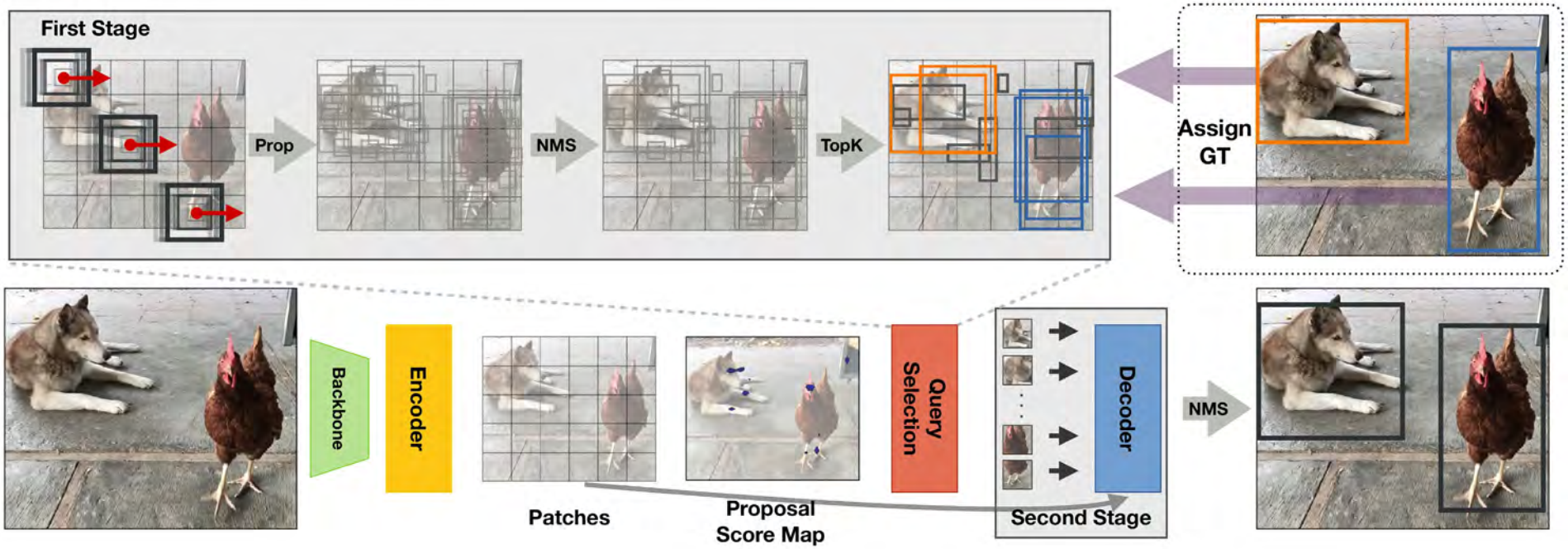
Redes de atención



Transformers

DETA tras DETR (12/2022)

No tiempo real



Transformers

DETA

COCO test

Method	Backbone	Extra Data	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
Deformable-DETR [40]	ResNeXt-101-DCN	-	50.1	69.7	54.6	30.6	52.8	65.6	4.6
EfficientDet-D7x [30]	EfficientNet-D7x-BiFPN	-	55.1	73.4	59.9	-	-	-	*6.5*
ScaledYOLOv4 [32]	CSPDarkNet-P7	-	55.4	73.3	60.7	38.1	59.5	67.4	*16*
CenterNet2 [38]	Res2Net-101-DCN-BiFPN	-	56.4	74.0	61.6	38.7	59.7	68.6	*5*
CopyPaste [10]	EfficientNet-B7	Objects365	56.0	-	-	-	-	-	-
HTC++ [3, 21]	Swin-L	-	57.7	-	-	-	-	-	-
$\mathcal{H}$-Deformable-DETR [13]	Swin-L	-	58.3	**77.1**	63.9	**39.8**	61.5	72.7	4.6
DINO [36]	Swin-L	-	**58.6**	76.9	64.1	39.4	61.6	73.2	2.7
Improved Deformable-DETR	Swin-L	-	56.6	75.6	61.9	38.8	60.4	73.5	4.3
DETA (Ours)	Swin-L	-	58.5	76.5	**64.4**	38.5	**62.6**	**73.8**	4.2
DINO [36]	Swin-L	Objects365	63.3	-	-	-	-	-	2.7
DETA (Ours)	Swin-L	Objects365	**63.5**	**80.4**	**70.2**	**46.1**	**66.9**	**76.9**	4.2

Transformers

GroundingDINO

Fusiona conceptos de:

- DINO, detector basado en Transformers
- GLIP, asocia palabras a elementos viduales

Adecuado para *open set*, nuevas categorías

Resultados COCO

Model	Backbone	Pre-Training Data	Zero-Shot 2017val	Fine-Tuning 2017val/test-dev
Faster R-CNN	RN50-FPN	-	-	40.2 / -
Faster R-CNN	RN101-FPN	-	-	42.0 / -
DyHead-T [5]	Swin-T	-	-	49.7 / -
DyHead-L [5]	Swin-L	-	-	58.4 / 58.7
DyHead-L [5]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [50]	Swin-L	O365,SS-COCO	-	60.7 / 61.3
DINO(Swin-L) [57]	Swin-L	O365	-	62.5 / -
DyHead-T† [5]	Swin-T	O365	43.6	53.3 / -
GLIP-T (B) [25]	Swin-T	O365	44.9	53.8 / -
GLIP-T (C) [25]	Swin-T	O365,GoldG	46.7	55.1 / -
GLIP-L [25]	Swin-L	FourODs,GoldG,Cap24M	49.8	60.8 / 61.0
DINO(Swin-T)† [57]	Swin-T	O365	46.2	56.9 / -
Grounding DINO T (Ours)	Swin-T	O365	46.7	56.9 / -
Grounding DINO T (Ours)	Swin-T	O365,GoldG	48.1	57.1 / -
Grounding DINO T (Ours)	Swin-T	O365,GoldG,Cap4M	48.4	57.2 / -
Grounding DINO L (Ours)	Swin-L	O365,OI [19],GoldG	**52.5**	**62.6** / **62.7** (**63.0** / **63.0**)*
Grounding DINO L (Ours)	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	**60.7**‡	**62.6** / -

Transformers

GroundingDINO
Entrada por *prompt*
Útil para anotar y entrenar detector, p.e. yolov8

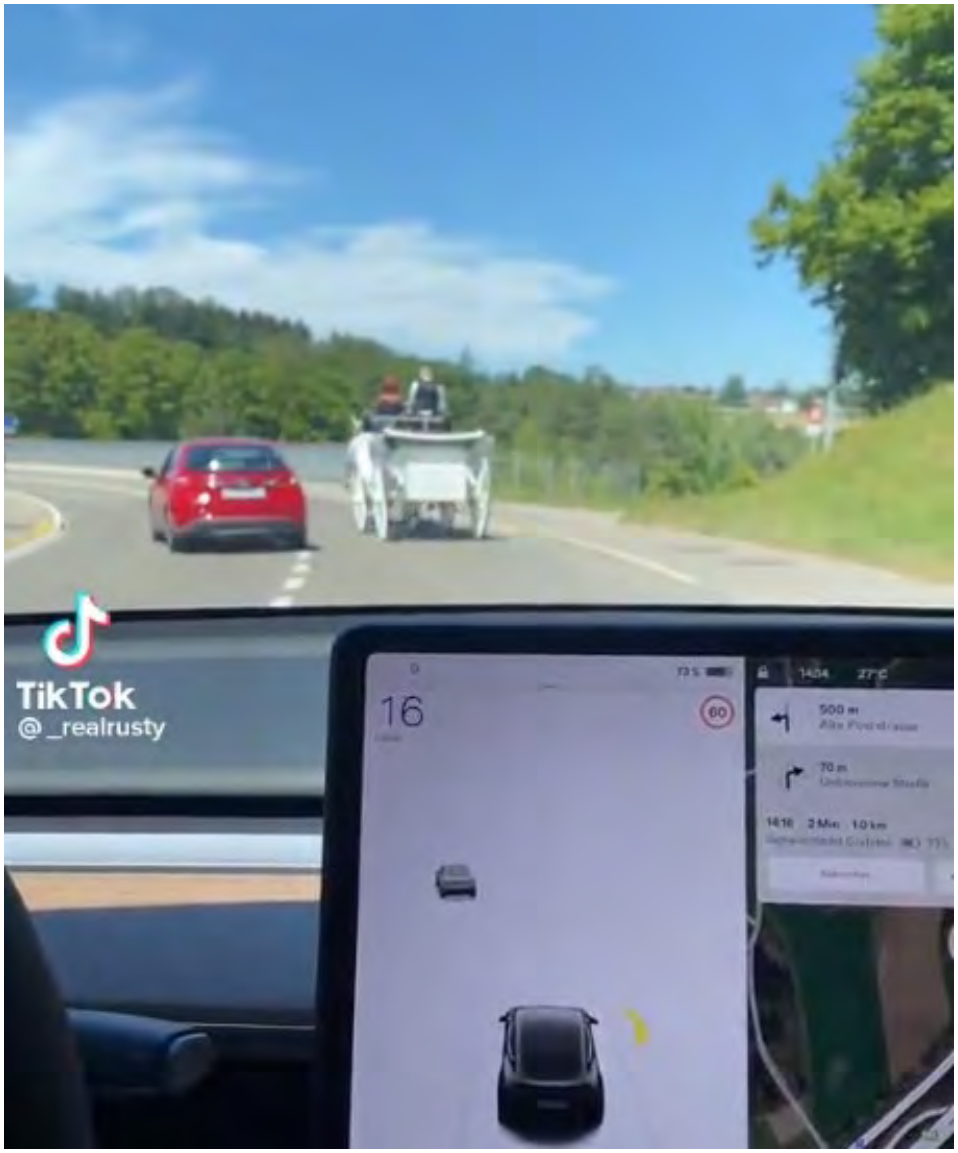
Ante el *prompt* “*window*”

Transformers

GroundingDINO
Entrada por *prompt*
Útil para anotar y entrenar detector, p.e. yolov8

Ante el *prompt* “*bird*”

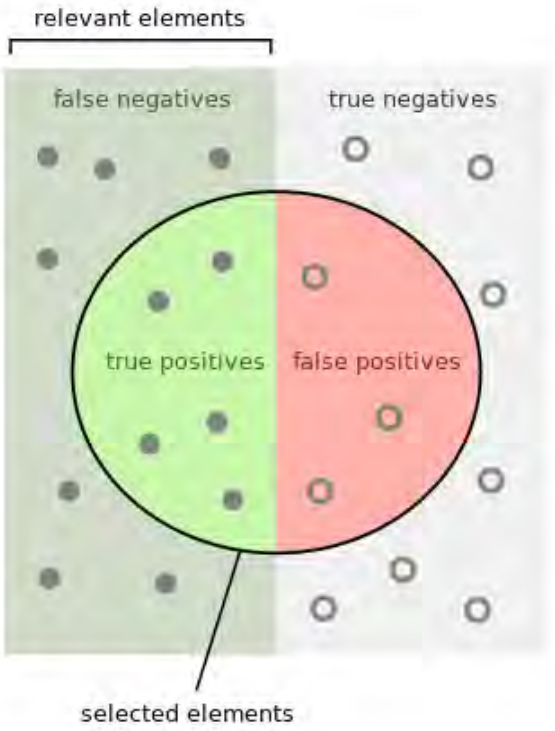
Detección

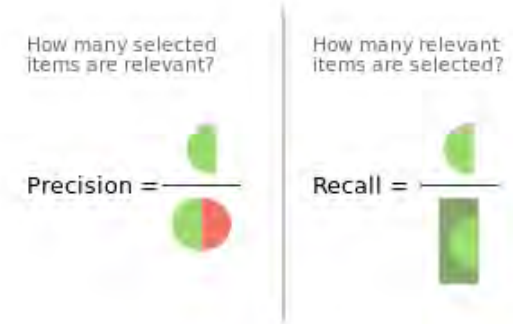


Evaluación del detector

Contabilizar aciertos y fallos

- Precisión
- Exhaustividad (*recall*)



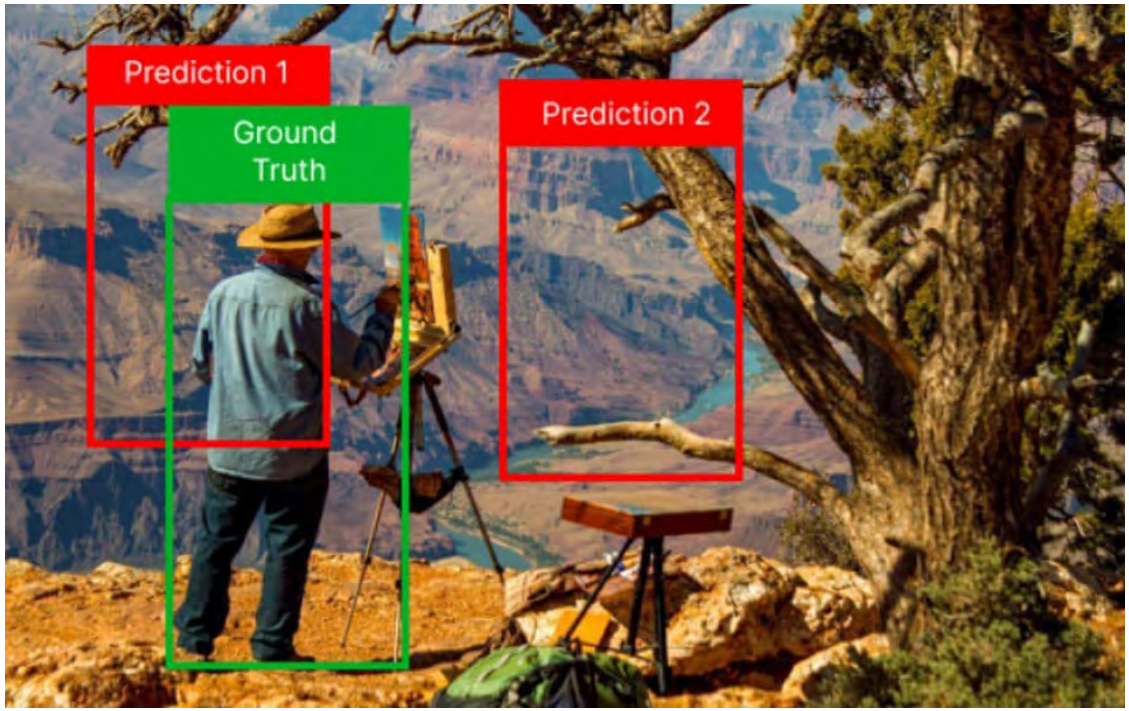


Evaluación del detector

Contabilizar aciertos y fallos

Anotaciones y detecciones

¿Cómo decidir si hay encaje de detección con anotación?



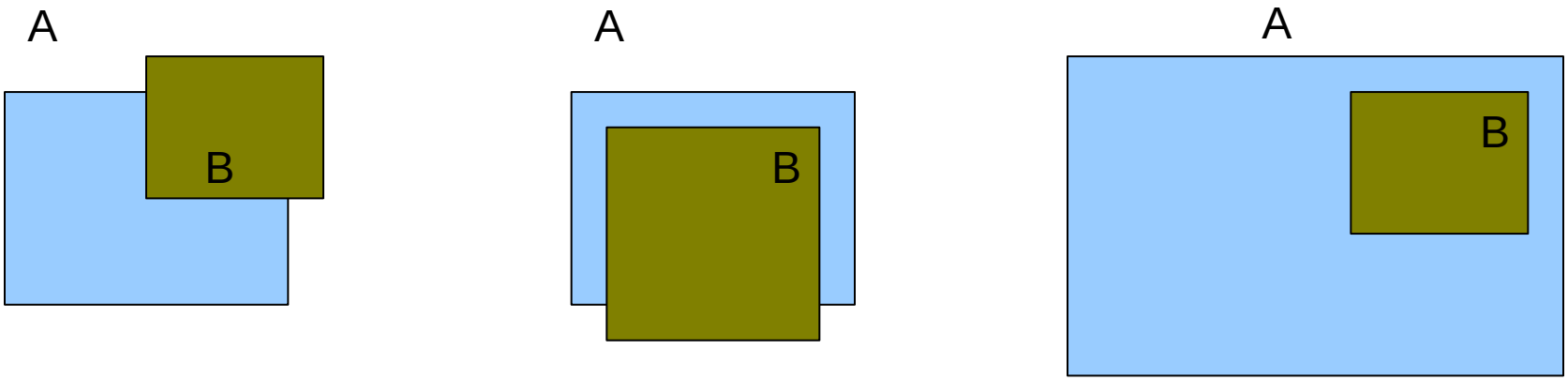
Evaluación del detector

Anotaciones y detecciones
¿Cómo decidir si hay encaje de detección con anotación?

IoU o índice de Jaccard

$$JI = \frac{A \cap B}{A \cup B}$$

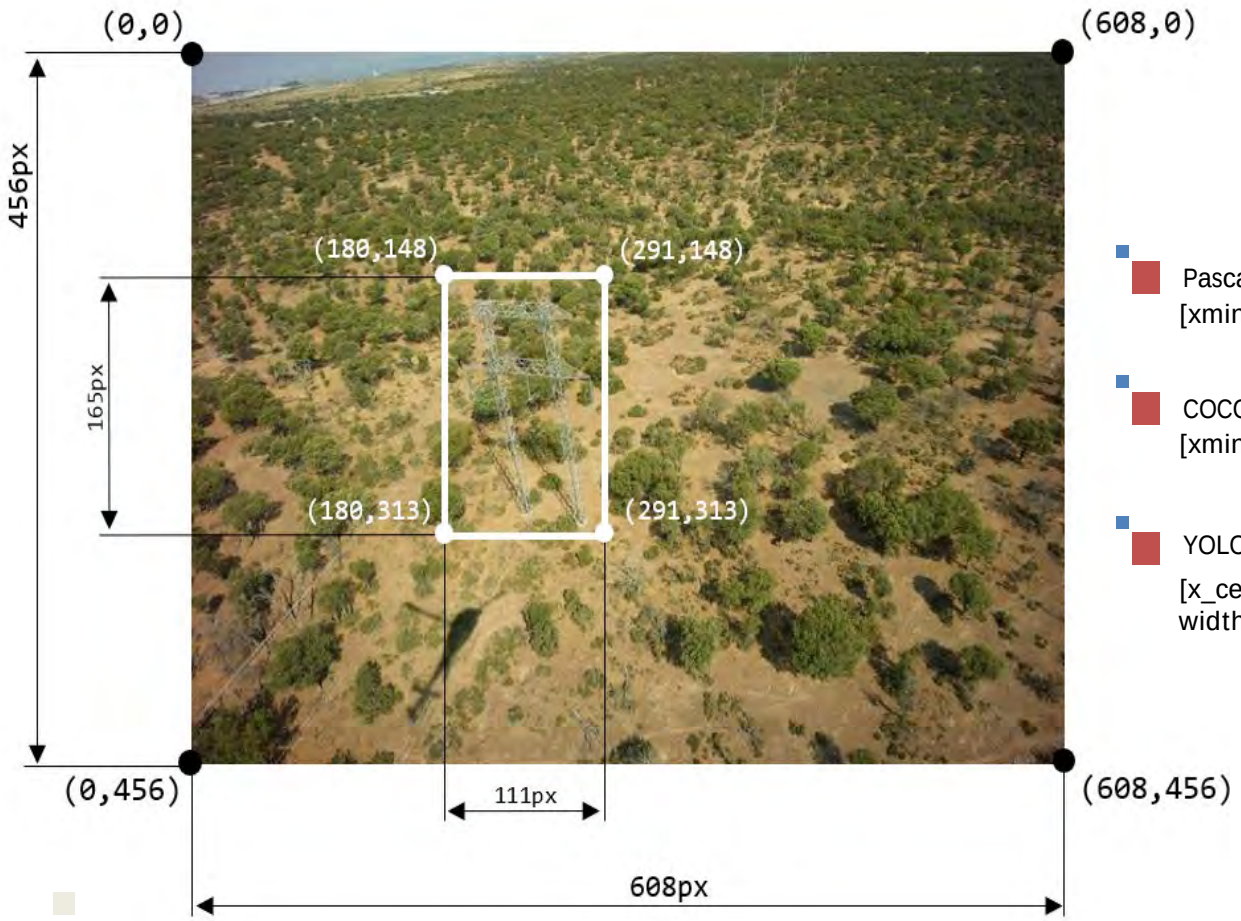
Umbral sobre JI para decidir si hay encaje o no
Valores menores, más encajes



Evaluación del detector

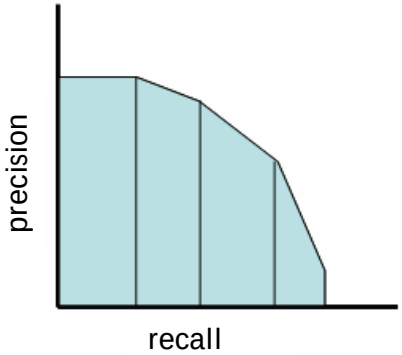
Anotaciones

Formatos diferentes



- Pascal VOC
  [xmin,ymin,xmax,ymax]
- COCO
  [xmin,ymin,width,height]
- YOLO (normalizadas)
  [x_center,y_center, width,height]

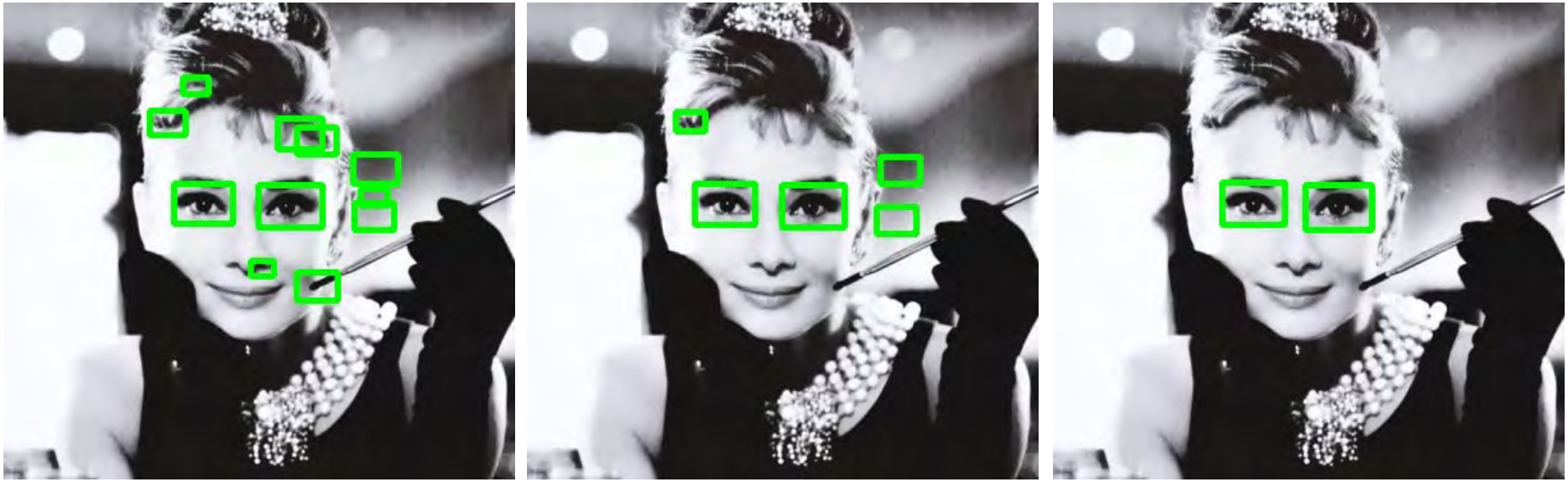
Evaluación del detector



Variar umbral de JI varían precisión y *recall*

Parámetros detector también hacen variar detecciones

$$precisión\ o\ PPV = \frac{TP}{TP + FP} = \frac{TP}{P} \qquad recall/sensitivity\ o\ TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$



Evaluación del detector

Average precision (AP)

Calcula el promedio curva *precisión/recall* de una clase

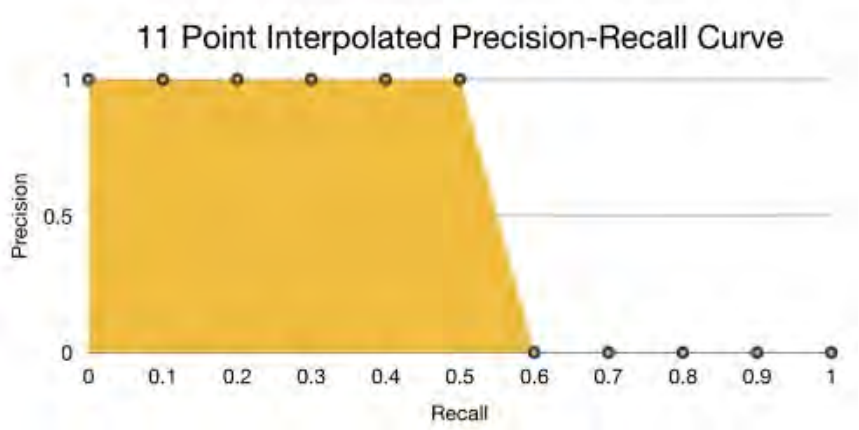
Puede interpolarse para eliminar outliers

Si la curva tiene *n* medidas:

$$AP = \frac{1}{n}\sum_{recall} precision(recall)$$

$$precisión = \frac{TP}{TP + FP}$$

$$exhaustividad = \frac{TP}{TP + FN}$$



Evaluación del detector

Mean average precisión (mAP)

AP, promedio de detección de una clase

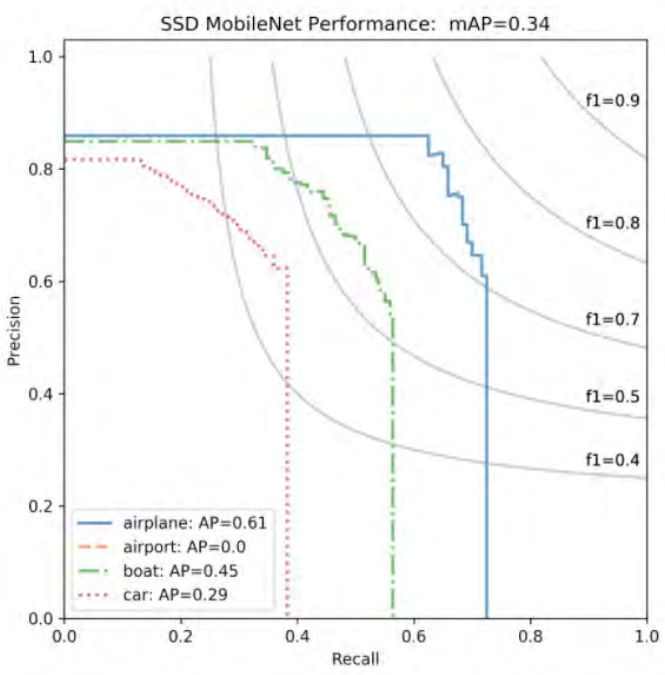
Para *nc* clases, el promedio de las respectivas APs

Con una única clase, mAP=AP

$$AP = \frac{1}{n}\sum_{recall} precision(recall)$$

Más de una clase

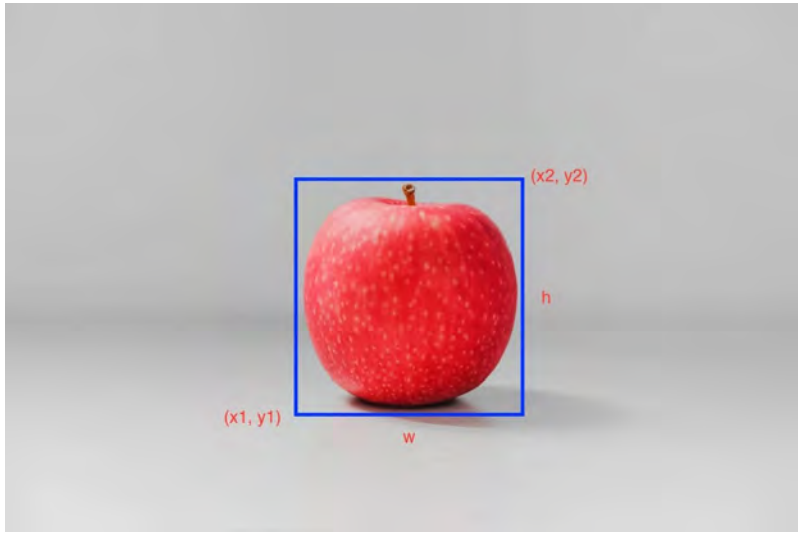
$$mAP = \frac{1}{nc}\sum_{clases} AP$$

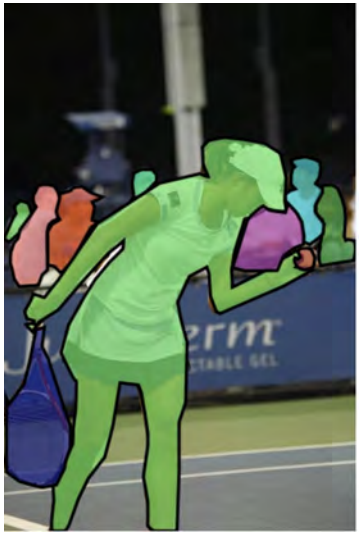


Evaluación del detector

¿Qué se anota?

- Contenedores
- Máscaras





Evaluación del detector

PASCAL Visual Object Challenge (PASCAL VOC 2009)

Veinte categorías

>500k imágenes de flickr

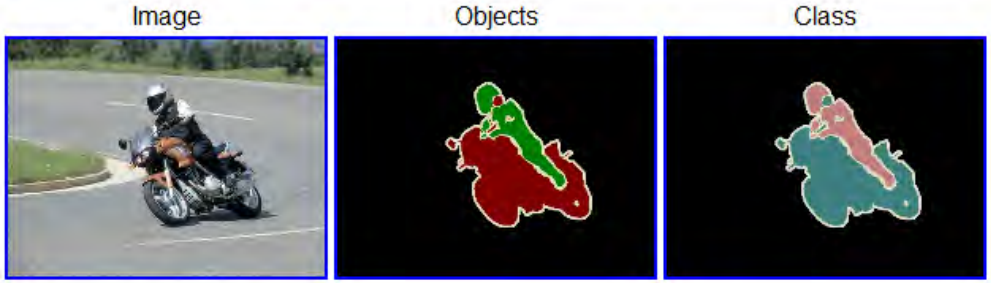
Anotación (web) tras selección aleatoria

Entrenamiento/validación

- 11,530 imágenes
- 27,450 contenedores
- 6,929 segmentaciones

Tareas:

- ¿Hay una moto en la imagen?
- Encuadra las motos de la imagen
- Segmenta las motos de la imagen



Evaluación del detector

COCO (2014)

Ochenta categorías

330k imágenes

1.5 millones de anotaciones

250k personas

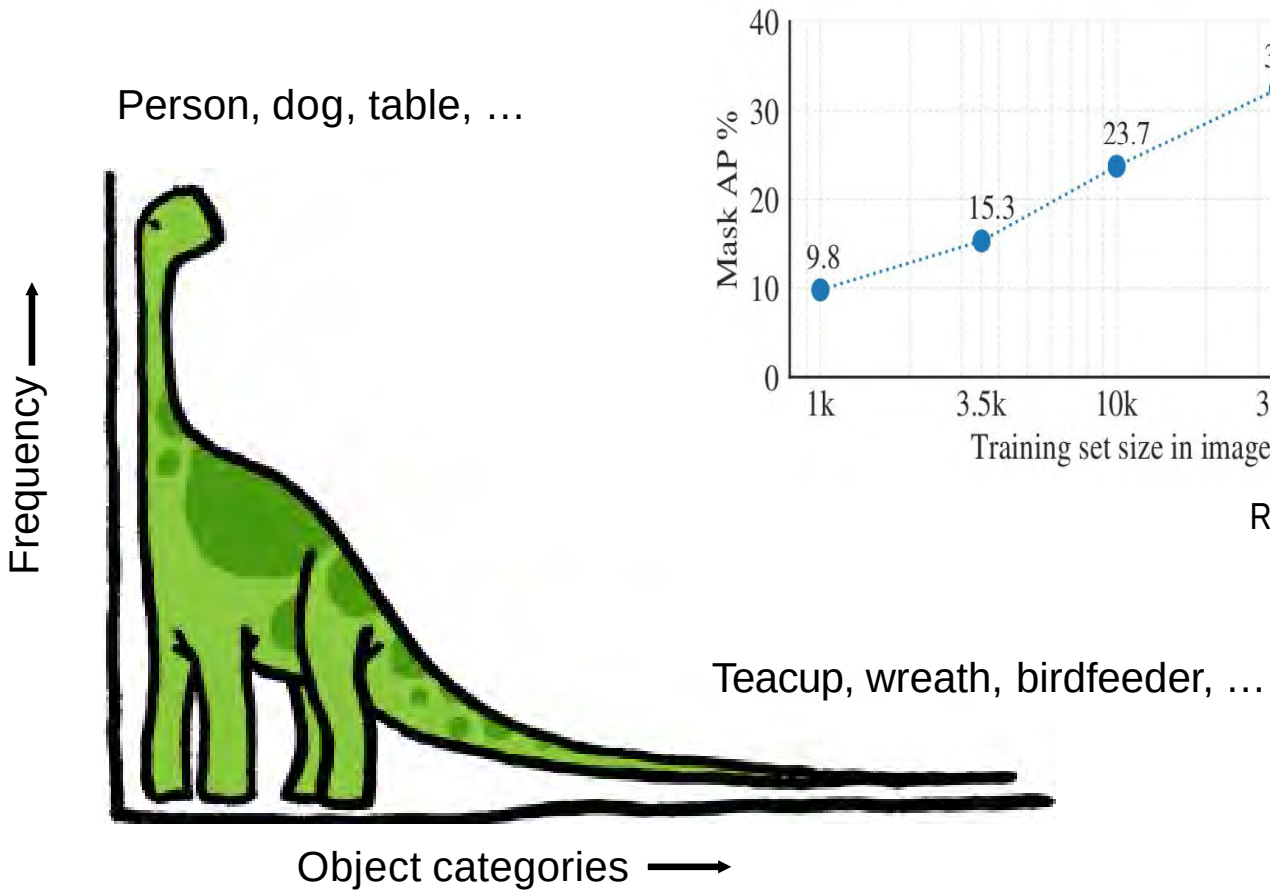
Evaluación del detector

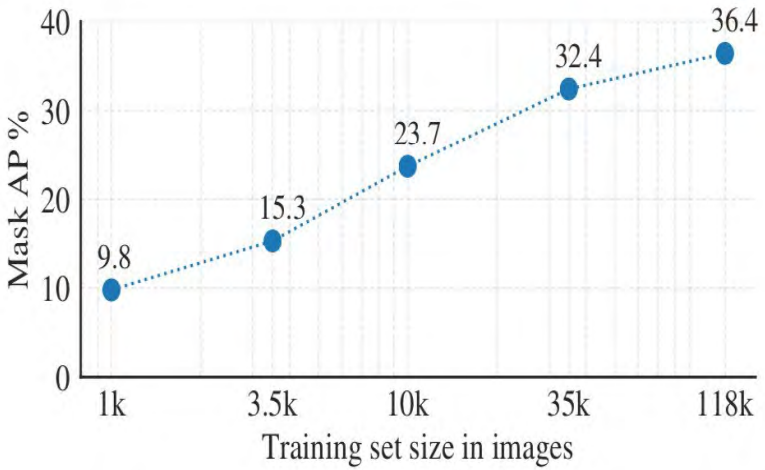
COCO

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast [6]	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
Fast [6]	07+12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster [2]	07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
Faster [2]	07+12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Faster [2]	07+12+COCO	78.8	84.3	82.0	77.7	68.9	65.7	88.1	88.4	88.9	63.6	86.3	70.8	85.9	87.6	80.1	82.3	53.6	80.4	75.8	86.6	78.9
SSD300	07	68.0	73.4	77.5	64.1	59.0	38.9	75.2	80.8	78.5	46.0	67.8	69.2	76.6	82.1	77.0	72.5	41.2	64.2	69.1	78.0	68.5
SSD300	07+12	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD300	07+12+COCO	79.6	80.9	86.3	79.0	**76.2**	57.6	87.3	88.2	88.6	60.5	85.4	**76.7**	**87.5**	**89.2**	84.5	81.4	55.0	81.9	**81.5**	85.9	78.9
SSD512	07	71.6	75.1	81.4	69.8	60.8	46.3	82.6	84.7	84.1	48.5	75.0	67.4	82.3	83.9	79.4	76.6	44.9	69.9	69.1	78.1	71.8
SSD512	07+12	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
SSD512	07+12+COCO	**81.6**	**86.6**	**88.3**	**82.4**	76.0	**66.3**	**88.6**	**88.9**	**89.1**	**65.1**	**88.4**	73.6	86.5	88.9	**85.3**	**84.6**	**59.1**	**85.0**	80.4	**87.4**	**81.2**

Evaluación del detector

Numero de categorías ¿cuántas?
Anotación (muchas para mejorar)





R. Girshick

Evaluación del detector

LVIS (*Large Vocabulary Instance Segmentation*)

>1200 categorías

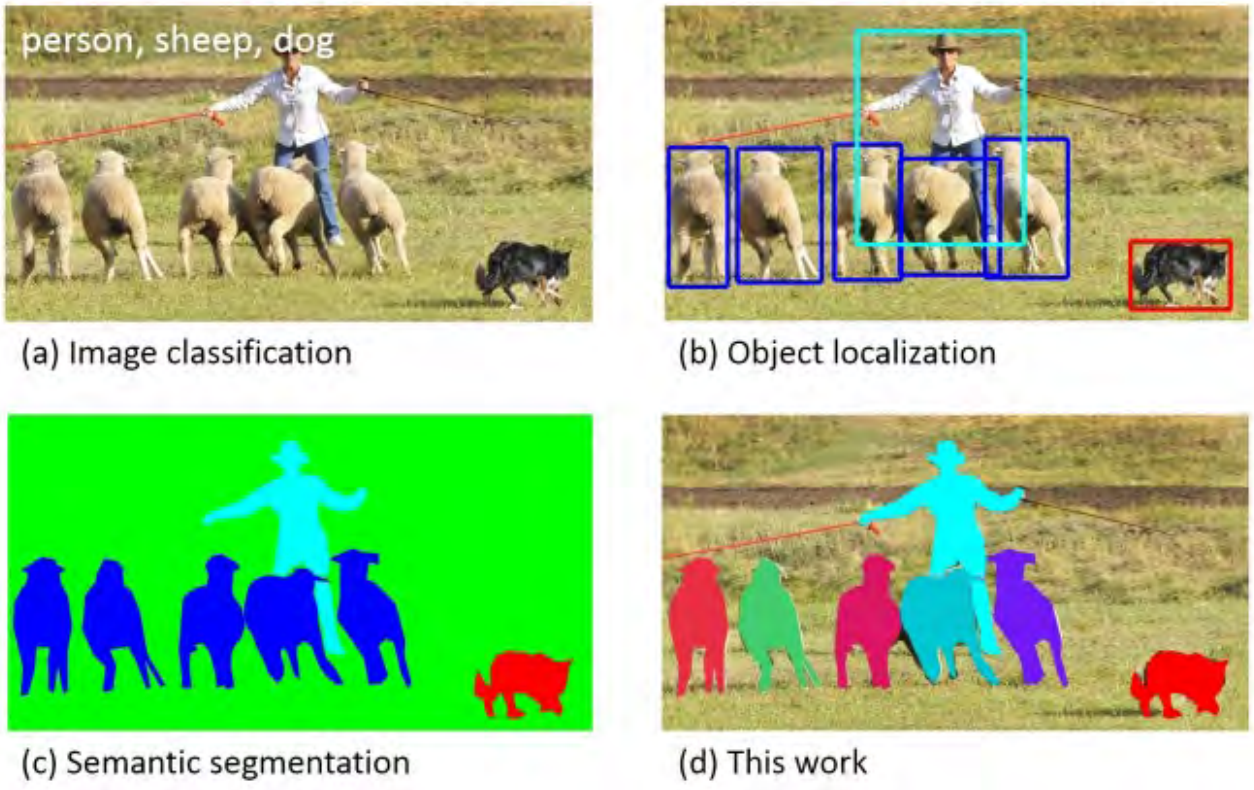
>2 millones de máscaras anotadas

Detección

Object recognition		
Oxford buildings dataset	Pictures of buildings	5,062 images
https://www.robots.ox.ac.uk/~vgg/data/oxbuildings		Philbin, Chum *et al.* (2007)
INRIA Holidays	Holiday scenes	1,491 images
https://lear.inrialpes.fr/people/jegou/data.php		Jégou, Douze, and Schmid (2008)
PASCAL	Segmentations, boxes	11k images (2.9k with segmentations)
http://host.robots.ox.ac.uk/pascal/VOC		Everingham, Eslami *et al.* (2015)
ImageNet	Complete images	21k (WordNet) classes, 14M images
https://www.image-net.org		Deng, Dong *et al.* (2009)
Fashion MNIST	Complete images	70k fashion products
https://github.com/zalandoresearch/fashion-mnist		Xiao, Rasul, and Vollgraf (2017)
Object detection and segmentation		
Caltech Pedestrian Dataset	Bounding boxes	Pedestrians
http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians		Dollár, Wojek *et al.* (2009)
MSR Cambridge	Per-pixel segmentations	23 classes
https://www.microsoft.com/en-us/research/project/image-understanding		Shotton, Winn *et al.* (2009)
LabelMe dataset	Polygonal boundaries	>500 categories
http://labelme.csail.mit.edu		Russell, Torralba *et al.* (2008)
Microsoft COCO	Segmentations, boxes	330k images
https://cocodataset.org		Lin, Maire *et al.* (2014)
Cityscapes	Polygonal boundaries	30 classes, 25,000 images
https://www.cityscapes-dataset.com		Cordts, Omran *et al.* (2016)
Broden	Segmentation masks	A variety of visual concepts
http://netdissect.csail.mit.edu		Bau, Zhou *et al.* (2017)
Broden+	Segmentation masks	A variety of visual concepts
https://github.com/CSAILVision/unifiedparsing		Xiao, Liu *et al.* (2018)
LVIS	Instance segmentations	1,000 categories, 2.2M images
https://www.lvisdataset.org		Gupta, Dollár, and Girshick (2019)
Open Images	Segs., relationships	478k images, 3M relationships
https://g.co/dataset/openimages		Kuznetsova, Rom *et al.* (2020)

Segmentación semántica

Predecir máscara de segmentación de cada objeto



Lin, Tsung-Yi, et al. Microsoft coco: Common objects in context. Computer Vision–ECCV 2014

Segmentación semántica

Problemas:

- Semántica: etiquetado de clase por píxel
- De ejemplares: delinear objetos relevantes
- *Panoptic*: etiquetar objetos y resto
- Estimación pose: píxeles de personas y sus partes

Primeras propuestas etiqueta el píxel:

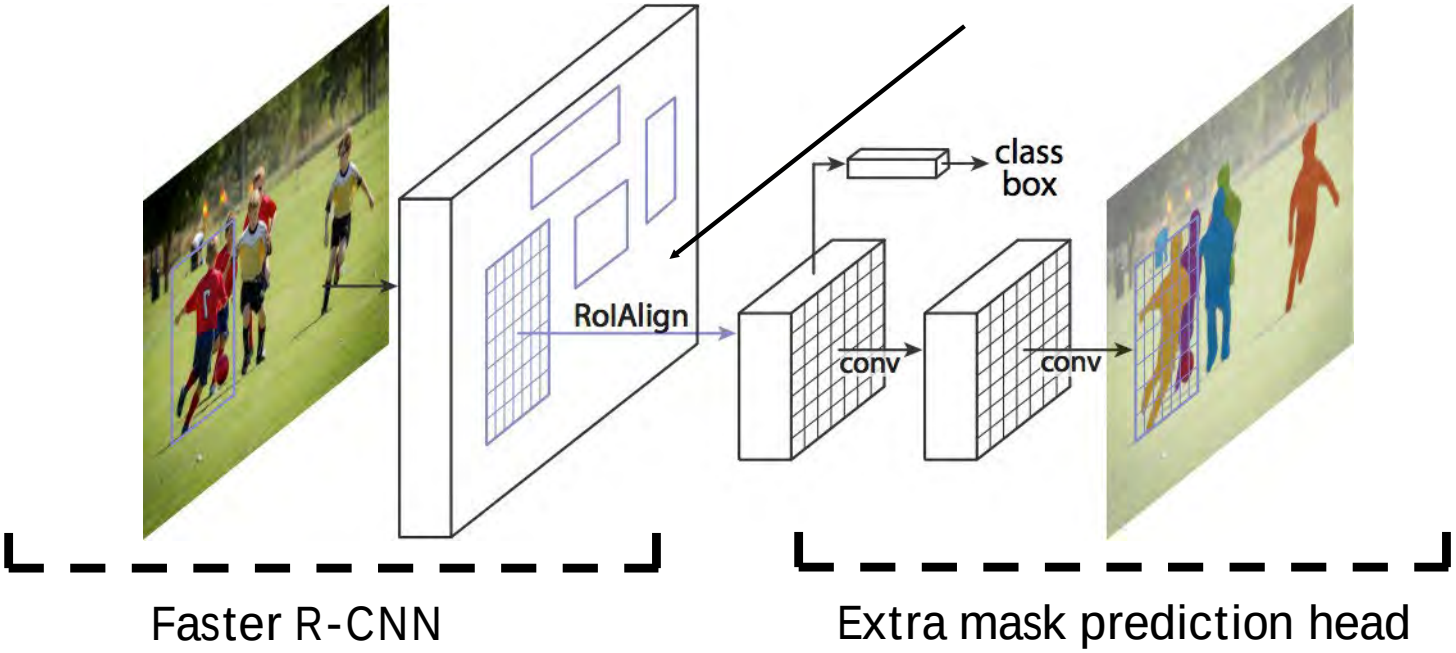
- Color
- Localización y contexto

Actuales usan pirámides y CNN

Evaluados en Pascal VOC, COCO, Cityscapes, ADE20K

Ejemplares (*instance*)

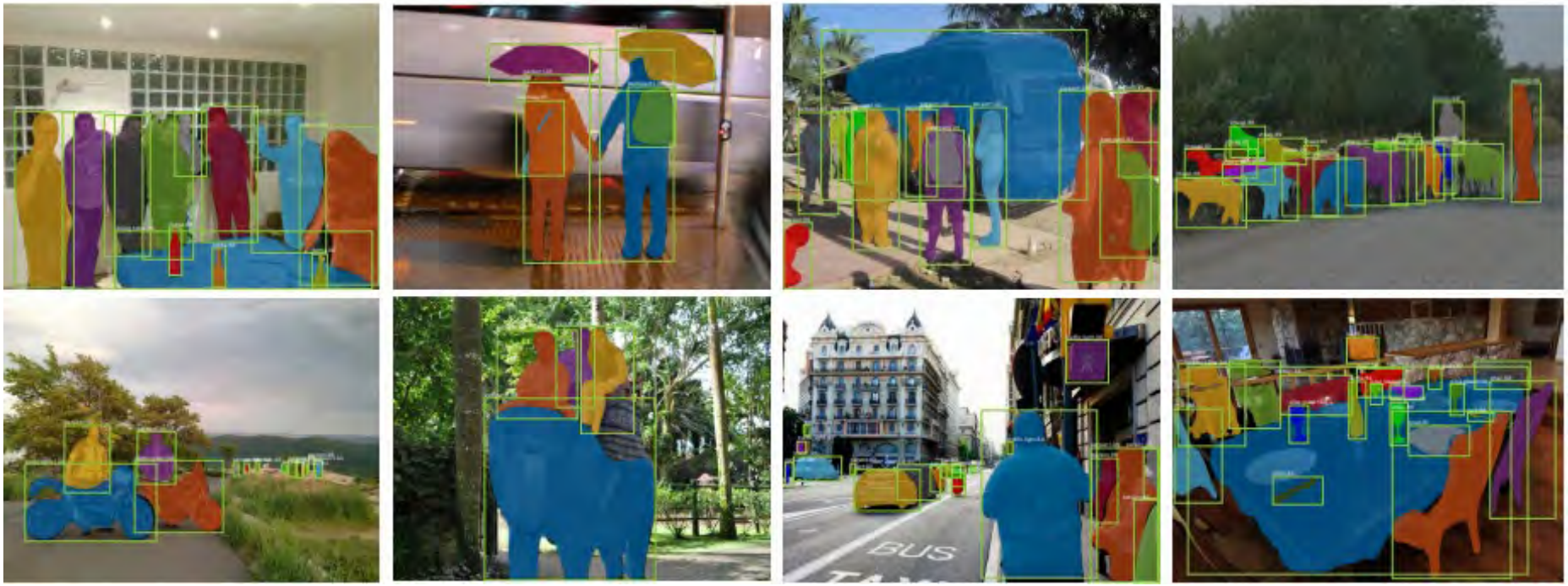
Máscaras de objetos relevantes
Hito en 2017 con Mask R-CNN
Reutiliza propuesta de regiones Faster R-CNN
Mejoras posteriores en arquitectura *backbone*



K. He et al., “Mask R-CNN”, ICCV 2017

Ejemplares (*instance*)

Mask R-CNN



Ejemplares (*instance*)

U-Net

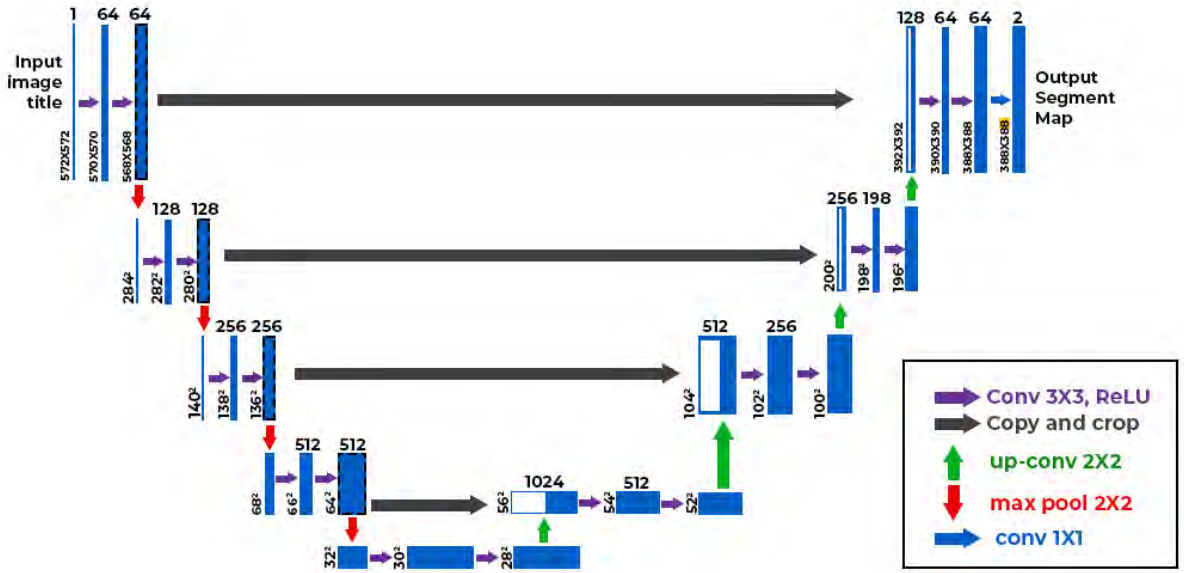
Basada en CNN

Origen biomédico, eficaz conpocos datos disponibles

Estructura simétrica en U, codificador-decodificador

Conexiones de salto preservan información espacial

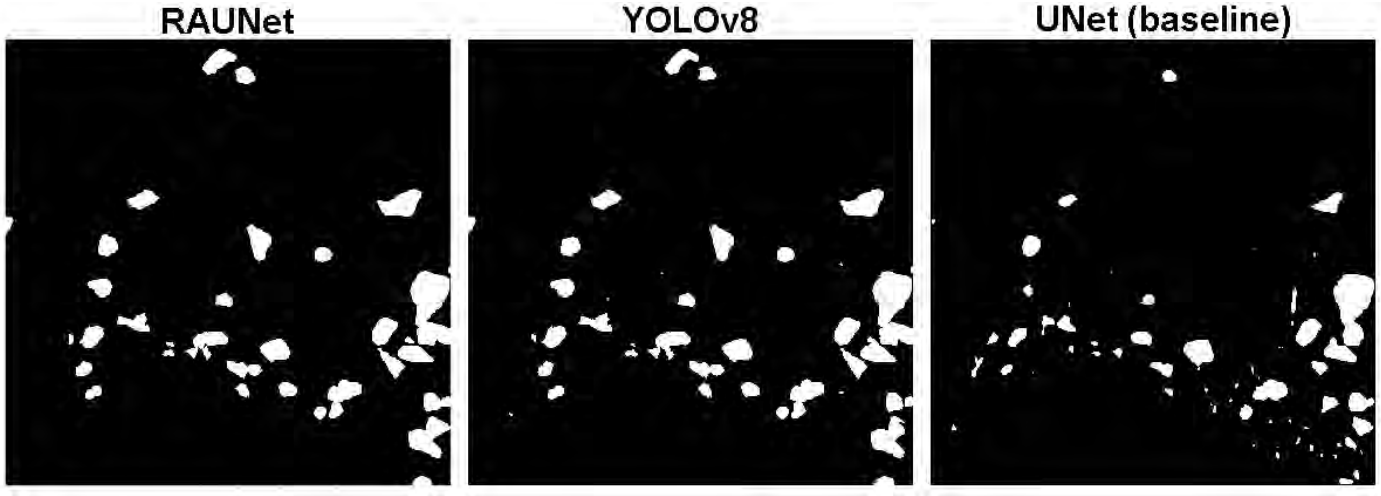
Segmentación a nivel de píxel



Ejemplares (*instance*)

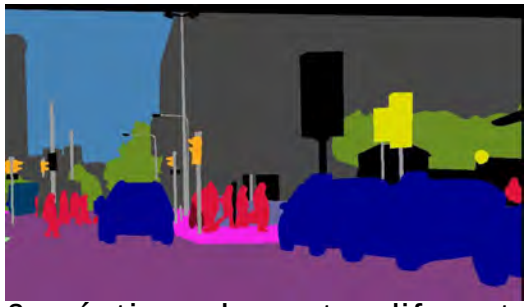
U-Net

Familia (UNet++, UNet 3+, ResUNet, RAUNet, ...)



Panóptica (*panoptic*)

Segmenta objetos, etiquetando el resto de la imagen



Semántica: elementos diferentes

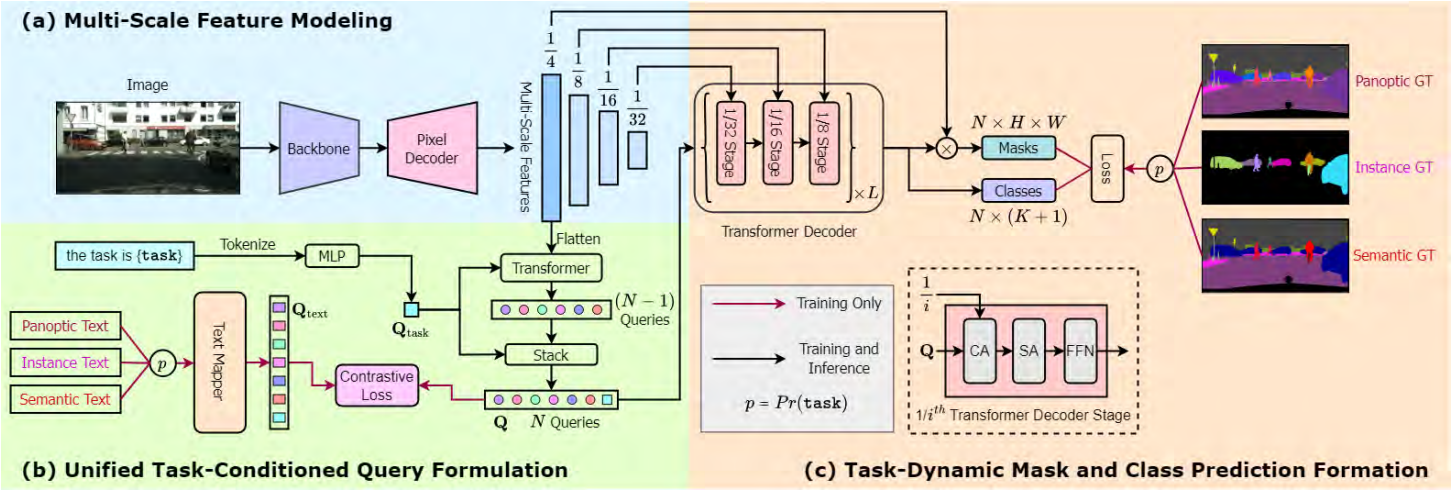
Ejemplares -> objetos

Panóptica -> segmentación por píxel y etiqueta objetos

A. Kirillov et al. Panoptic Segmentation, CVPR 2019

Panóptica (*panoptic*)

SAM



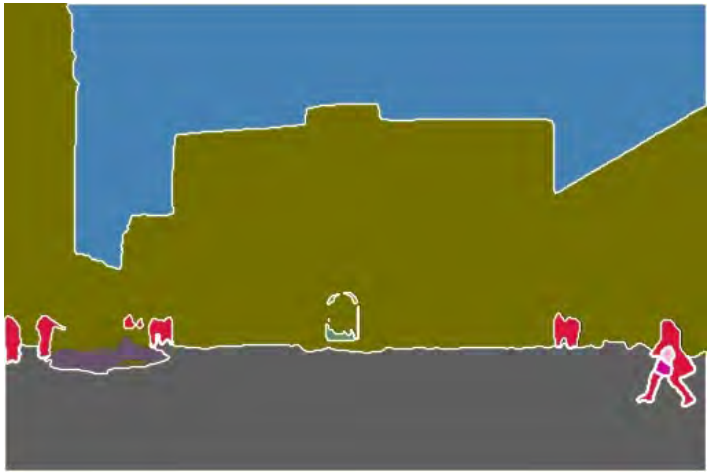
GroundingDINO+SAM

Grounded-SAM

Panóptica (*panoptic*)

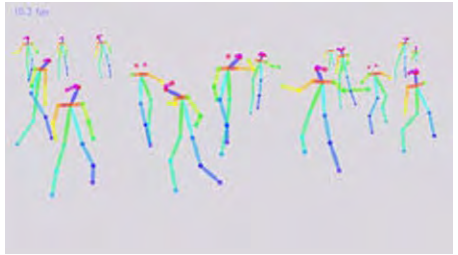
Oneformer

[Demo]



Estimación de pose

Determinar pose de personas y sus partes
DPM proporcionaba localización de partes



Estima localización del esqueleto en 2D:

- AlphaPose
- OpenPose
- …

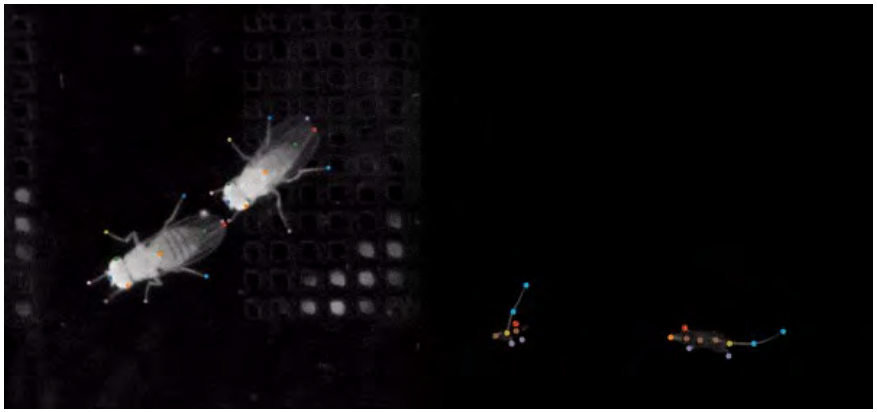
Proporciona etiquetado de partes por píxel:

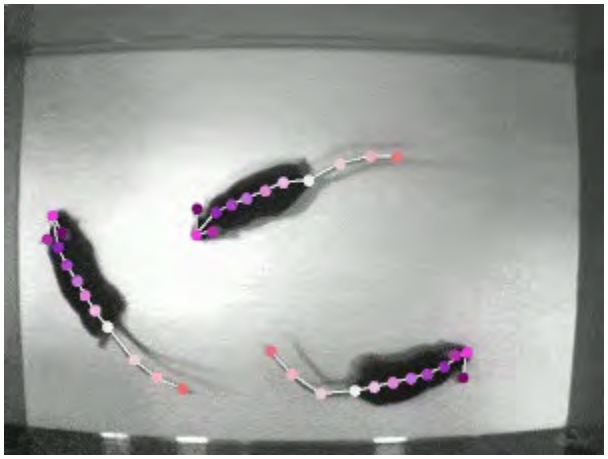
- DensePose (sobre R-CNN)
- …

MediaPipe

Estimación de pose

No sólo personas: SLEAP, DeepLabCut





Clasificación/Reconocimiento

Dos grandes grupos en reconocimiento de objetos:

- Ejemplar/Instancia (*instance recognition*)
  Localizar objetos conocidos rígidos
  Cambios de puntos vista, oclusiones

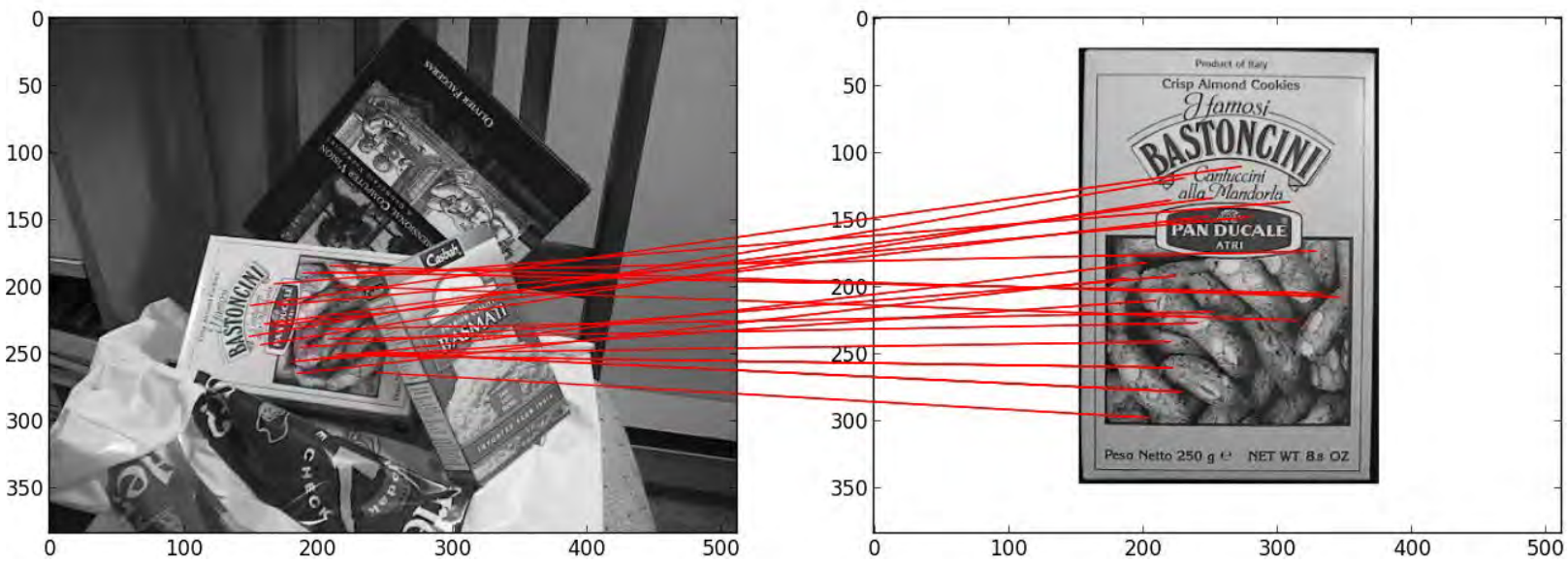
- Clases o categorías
  Reconocer cualquier ejemplar de una clase
  Problema más general

Ejemplares

Alineamiento geométrico

Encaje de puntos, bordes y líneas
Múltiples vistas y descomposición autovectores
Encaje de puntos de interés invariantes (p.e. SIFT)

Aplicaciones comerciales

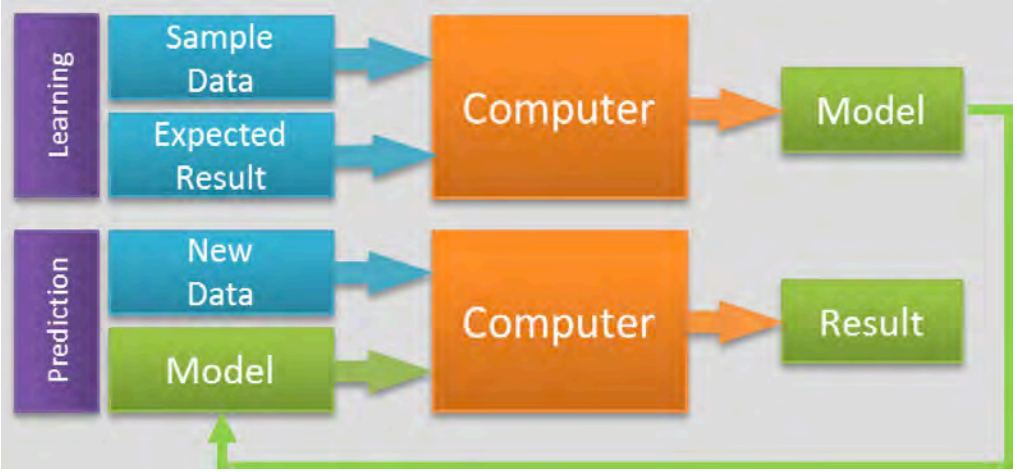


Enfoques

Gran desarrollo última década
Más conjuntos de datos disponibles

Tradicional (*hand-crafted*)

Aprendizaje profundo (DL)



Tradicionales

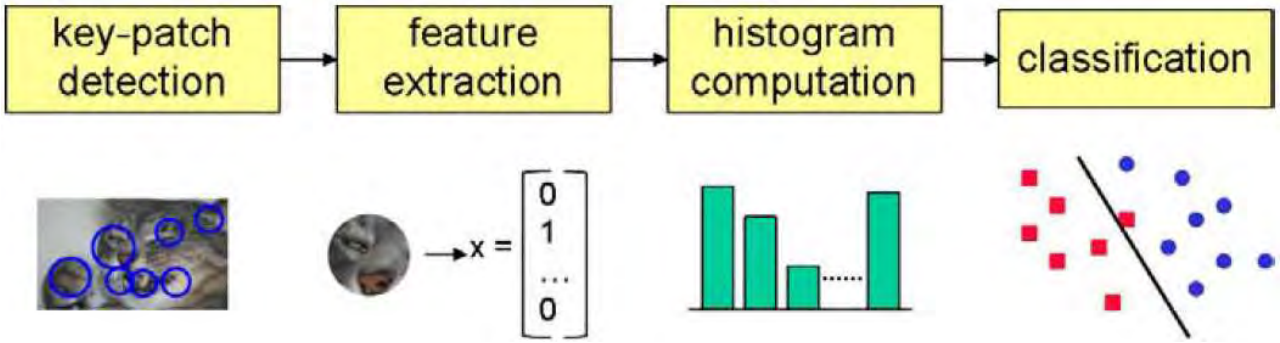
Bag of words/features/keypoints

Representación colección desordenada de descriptores

Histograma de *palabras* presentes en la imagen

Compara con entrenamiento

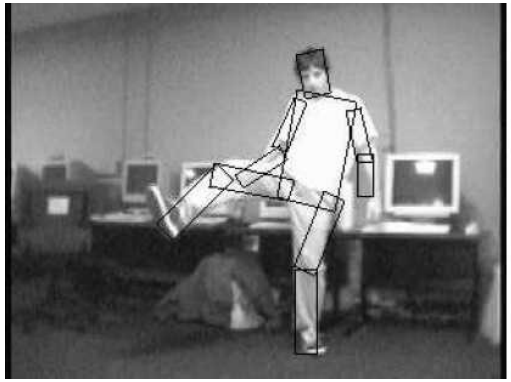
Sin verificación geométrica

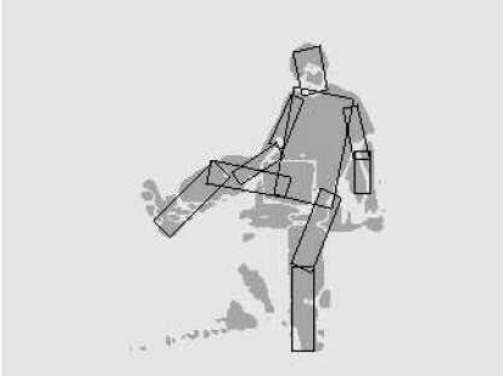


Csurka

Tradicionales

Basado en partes
Representan partes y relaciones geométricas
Importancia del contexto





Aprendizaje profundo

Desde AlexNet con ImageNet Large Scale Visual Recognition Challenge (ILSVRC) en 2012

Continuo crecimiento

Redes más eficientes

Mayores y mejores conjuntos de datos

Fine-grained categorization

Diferencias entre subcategorías sutiles (p.e. biología), pocas muestras

Zero-shot learning

Descripción de categorías no vistas antes

Aprendizaje profundo

Reconocimiento de caras → siguiente tema

Detección previa

Técnicas de evaluación experimental

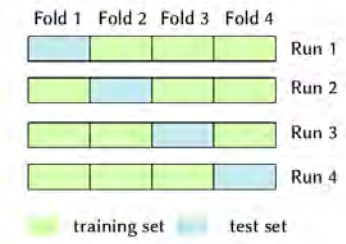
Hold-Out

El conjunto de datos se divide en dos subconjuntos, entrenamiento y test (p.e. 1/3)

Distribución escogida aleatoriamente

Las muestras de cada clase deben mantener la misma distribución en ambos conjuntos, es decir, balanceada

Técnicas de evaluación experimental



Validación cruzada k-fold

Particiona el conjunto de datos en un número de subconjuntos balanceados

Valida resultados, repitiendo el experimento:

- Divide muestras en k conjuntos disjuntos (*folds*)
- Inicializar la media del error a cero
- Para i=1 hasta k hacer
  - Entrena con muestras de todos los subconjuntos excepto el i
  - Usa subconjunto i como test para estimar el error
  - Añade a la media de error
- Fin para
- Calcula promedio de los errores

Usa todas las muestras para entrenamiento y test

Frecuente 5-fold y 10-fold

Técnicas de evaluación experimental

Validación aleatoria repetida

Divide las muestras de forma aleatoria en datos de entrenamiento y test

Repite entrenamiento y test, promediando errores

La ventaja sobre la validación k-fold es que no hay dependencia del número de iteraciones

Riesgo de no emplear todas las muestras para test, en incluso repetir algunas en su uso

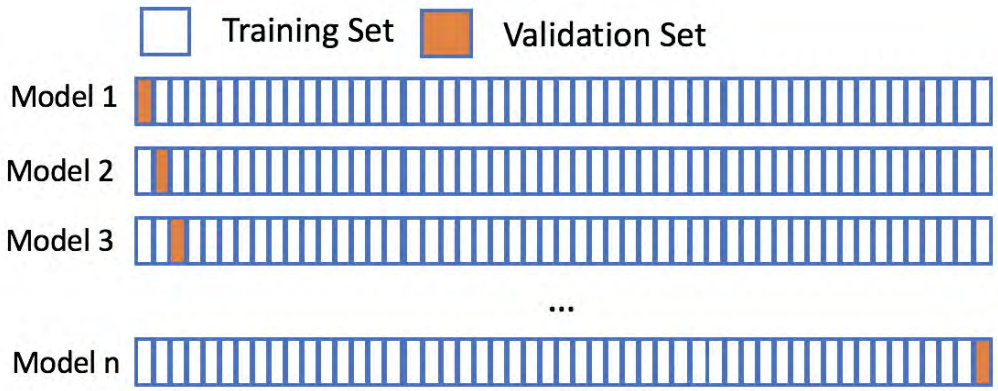
Técnicas de evaluación experimental

Leave-one-out cross-validation (LOOCV)

Utiliza una muestra como test y el resto para entrenamiento

Repite hasta emplear todas las muestras como test

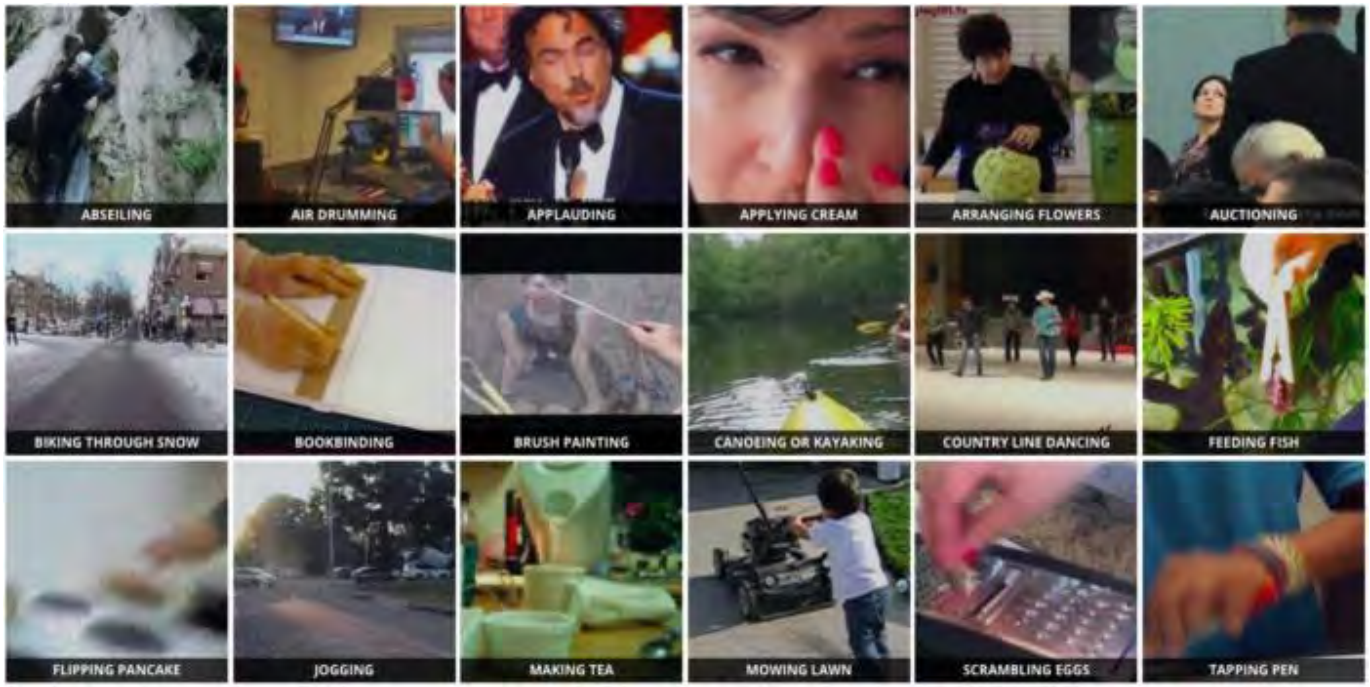
Similar a *k-fold* siendo *k* el número de muestras



Comprensión de vídeo

Relaciones entre personas y objetos

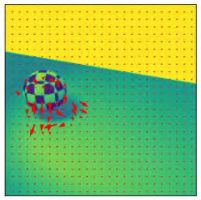
Human activity recognition (HAR)



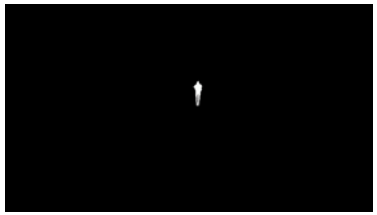
Comprensión de vídeo

Seguimiento (*tracking*)

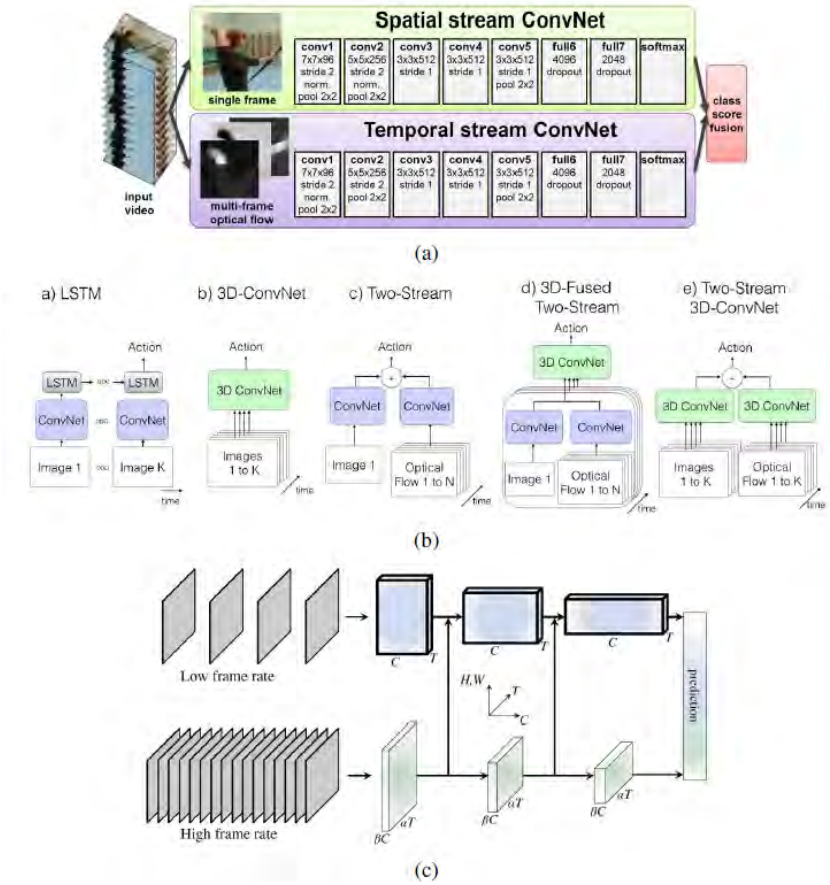
Flujo óptico



Silueta



Pose



Redes profundas:

- *Long short-term memory* (LSTM)
- 3D CNN
- etc.

Comprensión de vídeo

Conjuntos de datos

Name/URL	Metadata	Contents/Reference
Charades	Actions, objects, descriptions	9.8k videos
https://prior.allenai.org/projects/charades		Sigurdsson, Varol *et al.* (2016)
YouTube8M	Entities	4.8k visual entities, 8M videos
https://research.google.com/youtube8m		Abu-El-Haija, Kothari *et al.* (2016)
Kinetics	Action classes	700 action classes, 650k videos
https://deepmind.com/research/open-source/kinetics		Carreira and Zisserman (2017)
"Something-something"	Actions with objects	174 actions, 220k videos
https://20bn.com/datasets/something-something		Goyal, Kahou *et al.* (2017)
AVA	Actions	80 actions in 430 15-minute videos
https://research.google.com/ava		Gu, Sun *et al.* (2018)
EPIC-KITCHENS	Actions and objects	100 hours of egocentric videos
https://epic-kitchens.github.io		Damen, Doughty *et al.* (2018)

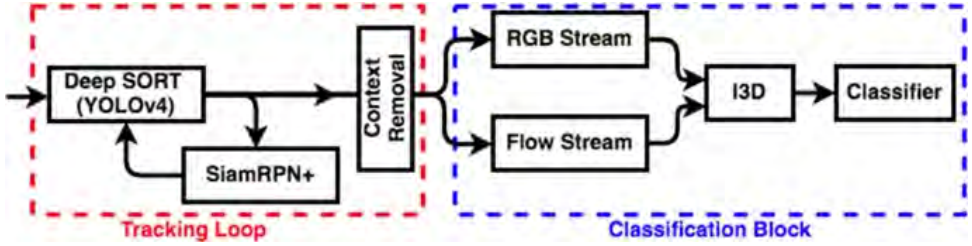
Comprensión de vídeo

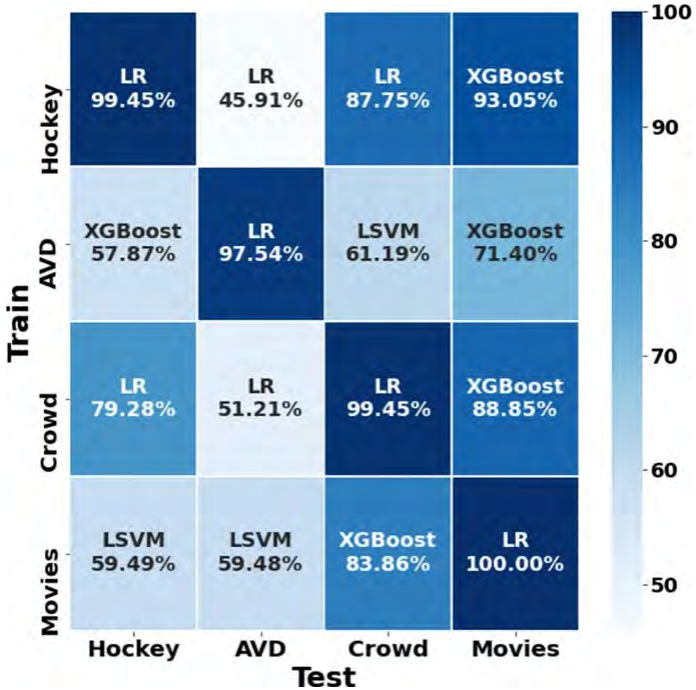
[Batallita]

RUNNERS/ILUSOS

Características I3D (flujo óptico y apariencia)

Detección de Violencia

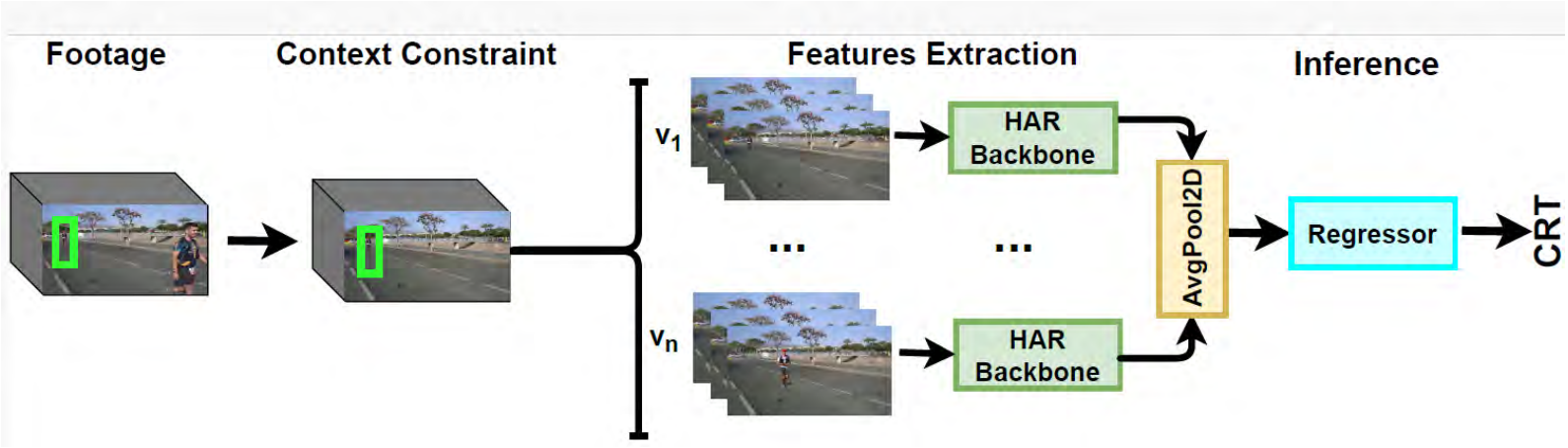




D. Freire et al. *Inflated 3D ConvNet context analysis for violence detection*. MVA 22

Comprensión de vídeo

[Batallita]
RUNNERS/ILUSOS
Características I3D, X3D
Estimación de tiempo de carrera 8-18 horas
9’ de error



D. Freire et al. *Towards cumulative race time regression in sports: I3D ConvNet transfer learning in ultra-distance running events*, ICPR 2022
D. Freire et al. *A Large-scale Analysis of Athletes’ Cumulative Race Time in Running Events*. ICIAP 23

Comprensión de vídeo

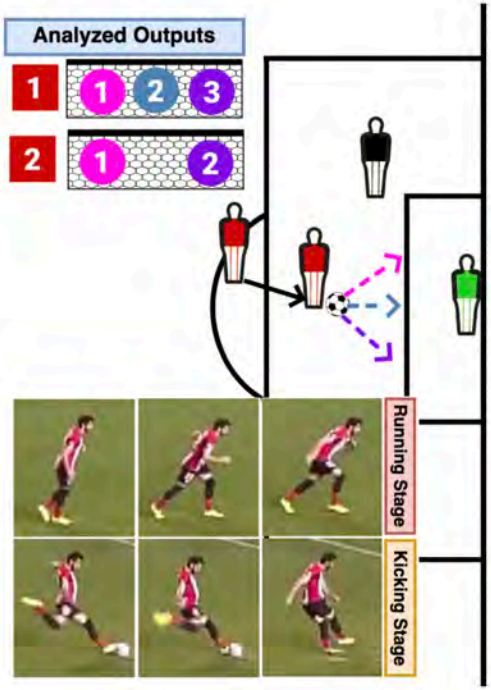
[Batallita]

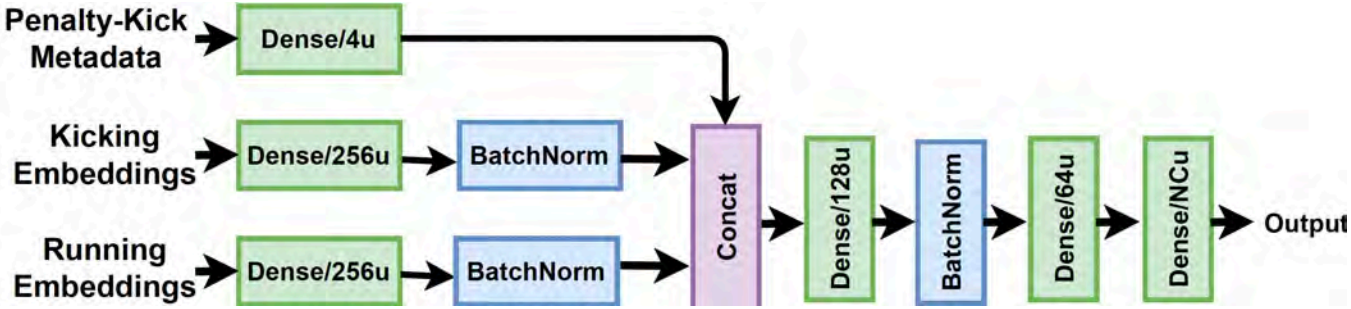
RUNNERS/ILUSOS

Lanzamientos de penalty

Predecir dirección balón

- 3 classes --> 51.9%
- 2 classes --> 63.9%





D. Freire et al. *Towards cumulative race time regression in sports: I3D ConvNet transfer learning in ultra-distance running events*, ICPR 2022

D. Freire et al. *A Large-scale Analysis of Athletes' Cumulative Race Time in Running Events*. ICIAP 23

Visión y lenguaje

Vision-Language Models (VLMs)

Propósito IA general

No basta construir modelos 3D o localizar imágenes

Integrar con otros componentes de IA como comprensión del habla y lenguaje, etc.

Combinar VC y procesamiento del lenguaje natural (NLP)

Avances en asistentes virtuales

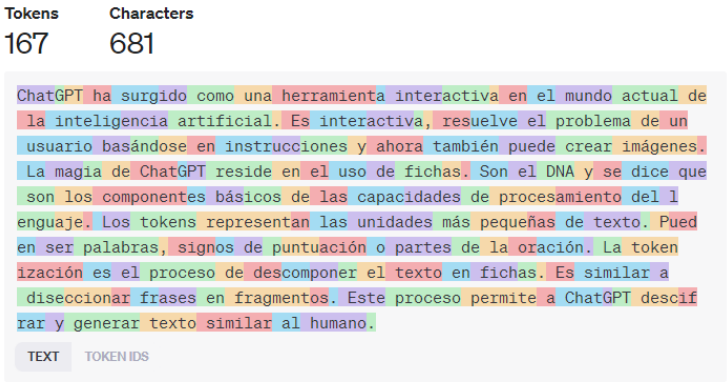
Visión y lenguaje

NLP

Codificar el texto en *k* símbolos

- Palabras, *k* muy alto
- Caracteres, cada palabra múltiples códigos
- *Byte-pair*, equilibrio entre ambas: GPT y CLIP

Códigos proyectado a *tokens*, entrada del *transformer*



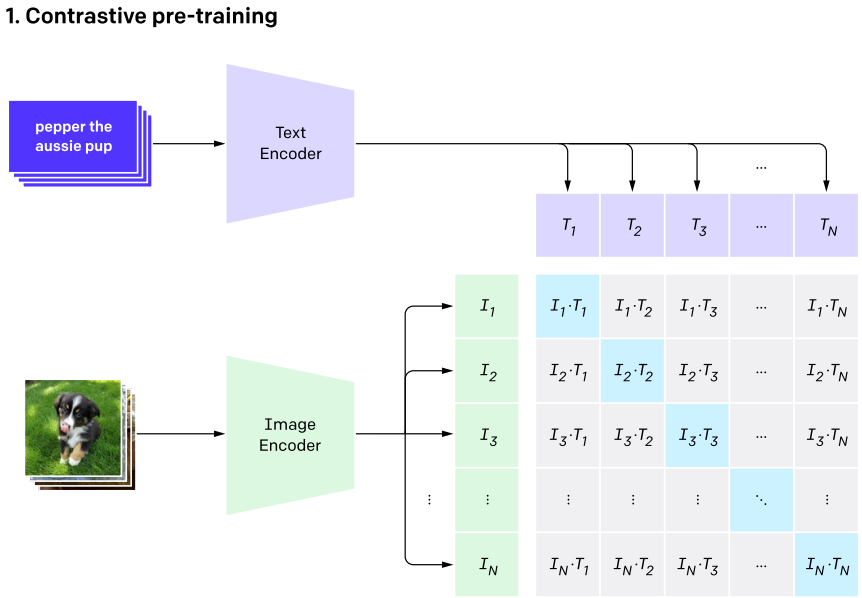
Visión y lenguaje

Conectar texto e imágenes

Esquema codificador/decodificador

- Codificador CNN detector o reconocedor de acciones
- Decodificador RNN o LSTM

CLIP empareja codificadores de texto e imagen



Visión y lenguaje

Image captioning
Subtitulado y descripción de imágenes

Vecino más cercano imágenes con similares subtítulos

Más reciente sistemas basados en atención

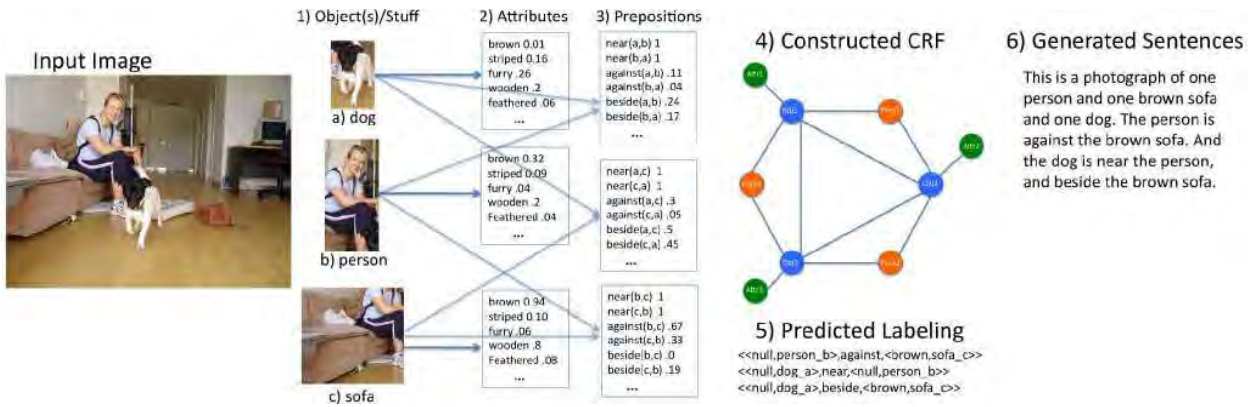
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

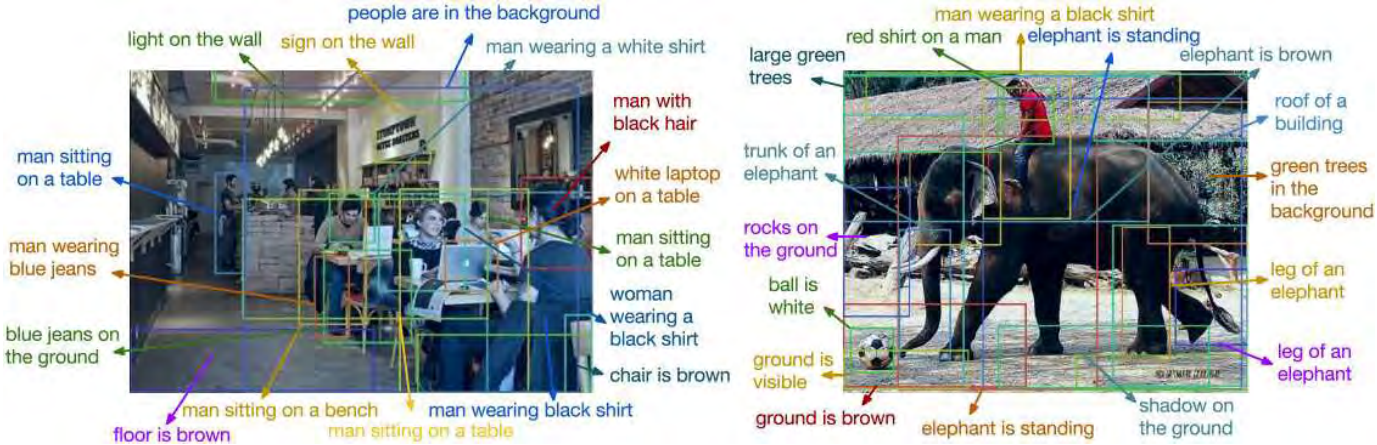
"two young girls are playing with lego toy."

Visión y lenguaje

BabyTalk



DenseCap



Visión y lenguaje

Text2image
Problema inverso

DALLE-3
Midjourney
Stable diffusion

Visión y lenguaje

[Batallita]
VQA PAR23

PAR CONTEST 2023

CAIP 2023

INTERNATIONAL CONTEST ON PEDESTRIAN ATTRIBUTE RECOGNITION

from partially annotat|

20th International Conference on Computer Analysis of Images and Patterns - CAIP 2023

CONTEST INSTRUCTIONS

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]
¿Por qué *pedestrians*?
Re-identificación:

- Corto plazo

Market1501

- Medio plazo (cambio moderado de indumentaria)

TGC20ReId

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Descripción semántica para reid?

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

'A young boy is seen in this surveillance image'

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

No apta para PAR23

'A young boy is seen in this surveillance image'

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

What color is the person shirt?

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

What color is the person shirt?

'Blue and white'

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

Tareas visión lenguaje

- *Image captioning*
- *Visual question answering* (*VQA*)

What color is the person shirt?

¡GUAU!

'Blue and white'

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]
Evaluando con corredores
PAR23 piedra de toque

Visión y lenguaje

[Batallita]
Evaluando con corredores
PAR23 piedra de toque fiabilidad

- Is the person in shorts? **yes**
- Is the person wearing trousers? **no**
- Does the person wear long socks? **yes**
- Does the person wear a backpack? **yes**
- What color is the backpack? **blue and white**
- What color is the person shirt? **blue**
- What color is the person trousers? **gray**
- Has the person shirt sleeves? **yes**
- What color are the person socks? **white**
- Is the person male or female? **male**
- Does the person wear sunglasses? **yes**
- Does the person wear a cap? **yes**
- Does the person wear a bib? **yes**
- What is the number on the bib? 2

Visión y lenguaje

[Batallita]
PAR23, cinco tareas
MIVIA dataset
- >93k entrenamiento
- >12k validación

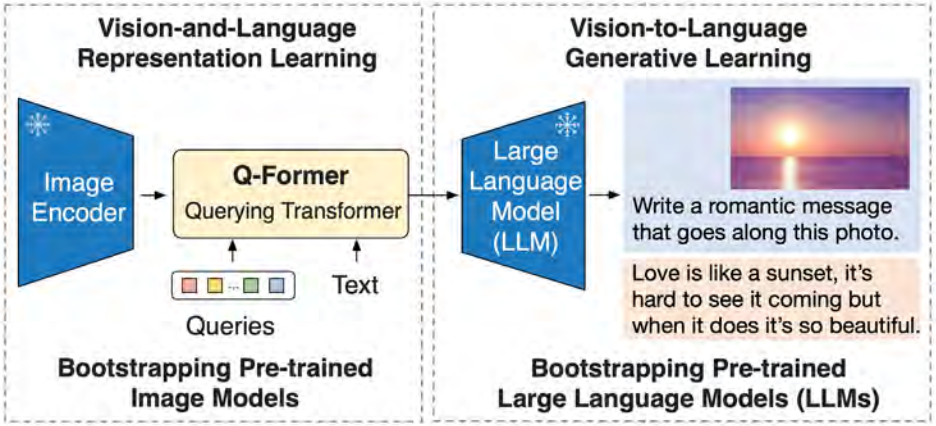
Etiquetas:
- Color vestimenta superior/inferior (11 colores anotados)
- Sexo
- Presencia de bolsos
- Presencia de sombrero

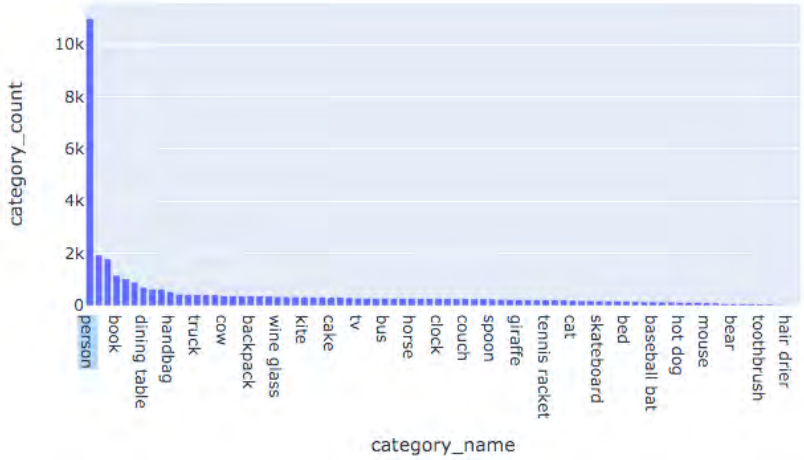
Color	Label
Black	1
Blue	2
Brown	3
Gray	4
Green	5
Orange	6
Pink	7
Purple	8
Red	9
White	10
Yellow	11

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]
BLIP-2
SOTA en varias tareas VyL
Modelos VQA
Entrenado con COCO
>204k imágenes
Muchas con personas
¿Bueno para describir?





M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]

BLIP-2

Preentrenado con VQAv2 (v2.0 2017)

- Balanceado
- ~1.1m preguntas
- ~11m respuestas

Goyal et al. (2019)

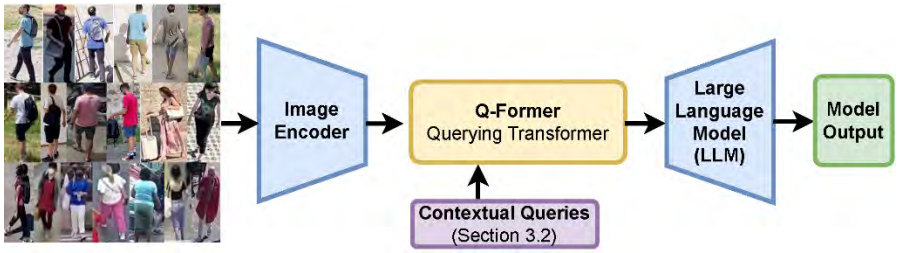
v2_OpenEnded_mscoco_train2014_q...

{"info": {"description": "This is v2.0 of the VQA dataset.", "url": "http://visualqa.org", "version": "2.0", "year": 2
"question_id": 458752000}, {"image_id": 458752, "question": "What position is this man playing?", "question_id": 45875
262146001}, {"image_id": 262146, "question": "What color is the persons headwear?", "question_id": 262146002}, {"image
"question": "Is there snow on the mountains?", "question_id": 393221001}, {"image_id": 393221, "question": "What is th
393223003}, {"image_id": 393224, "question": "Is the man smiling?", "question_id": 393224000}, {"image_id": 393224, "q
393224, "question": "Is the man wearing a plain tie?", "question_id": 393224005}, {"image_id": 524297, "question": "Ju
524297003}, {"image_id": 393227, "question": "Does the guy have a tattoo?", "question_id": 393227000}, {"image_id": 39
393227004}, {"image_id": 131084, "question": "What color is the jacket?", "question_id": 131084000}, {"image_id": 1310
"question_id": 131074001}, {"image_id": 131074, "question": "Are the walls done in a summery color?", "question_id": 1
{"image_id": 393230, "question": "How many people are in this image?", "question_id": 393230001}, {"image_id": 393230,
"question": "Did someone forget his luggage in the snow?", "question_id": 393230006}, {"image_id": 393230, "question"
"question_id": 131087000}, {"image_id": 131087, "question": "Is the dog wearing a collar?", "question_id": 131087001}

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]
Preguntas:



- *Is the person male or female?*
- *What color is the person's shirt?*
- *Does the person wear a jacket?*
- *What color is the person's jacket?*
- *What color is the person's trousers?*
- *Does the person wear a bag?*
- *Does the person wear a hat?*
- *Does the person wear a cap?*

M. Castrillón-Santana et al. *Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition.* CAIP 2023

Visión y lenguaje

[Batallita]
Color torso

Multicolor, se asigna el primero entre los 11
Color no presente en validación:

- Khaki
- Tan
- Plaid
- Camouflage
- … otros

Asignación manual o aleatoria

Predicted khaki

Predicted blue and white

Visión y lenguaje

[Batallita]
Color torso

Color de la chaqueta o camiseta

```
if person wears a jacket then
    color of upper body clothes = jacket color
else
    color of upper body clothes = shirt color
endif
```

Visión y lenguaje

[Batallita]

Resultados en validación, *zero-shot*

Task	Accuracy	Precision	Recall	F1
Upper color				
Lower color				
Gender				
Bag				
Hat				

$$Acc = \frac{\sum_1^K (p_i == g_i)}{K} \qquad precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN} \qquad f1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Visión y lenguaje

[Batallita]
Resultados en validación, *zero-shot*

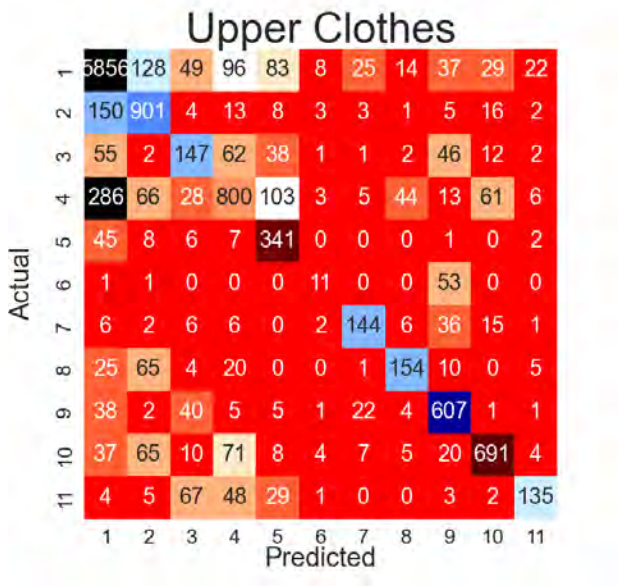
Task	Accuracy	Precision	Recall	F1
Upper color	0.805	0.805	0.805	0.801
Lower color	0.837	0.845	0.837	0.833
Gender	0.909	0.917	0.752	0.826
Bag	0.495	0.295	0.981	0.422
Hat	0.566	0.181	0.989	0.307

$$Acc = \frac{\sum_1^K (p_i == g_i)}{K} \qquad precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN} \qquad f1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Visión y lenguaje

[Batallita]

Matriz de confusión



Color	Label
Black	1
Blue	2
Brown	3
Gray	4
Green	5
Orange	6
Pink	7
Purple	8
Red	9
White	10
Yellow	11

Errores

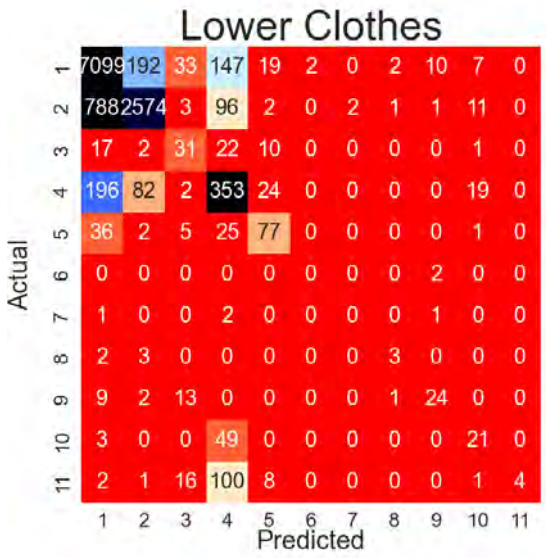
Annotated yellow
Predicted brown

Annotated black
Predicted blue and white

Visión y lenguaje

[Batallita]

Matriz de confusión



Color	Label
Black	1
Blue	2
Brown	3
Gray	4
Green	5
Orange	6
Pink	7
Purple	8
Red	9
White	10
Yellow	11

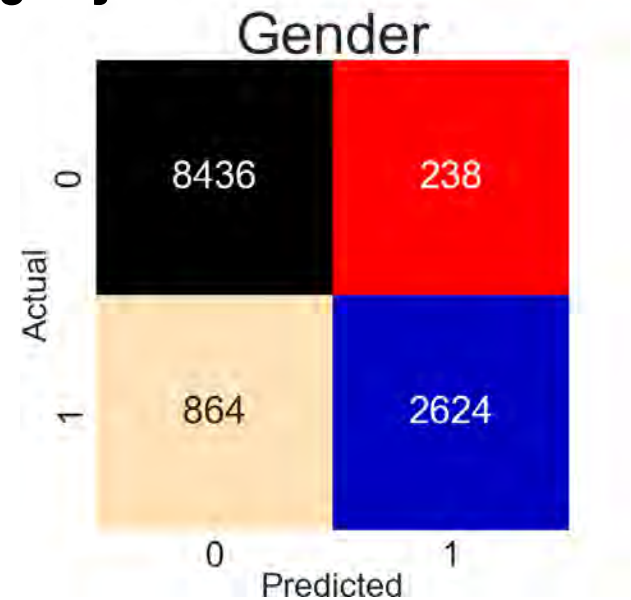
Errores

Annotated green
Predicted black

Annotated black
Predicted gray

Visión y lenguaje

[Batallita]
Matriz de confusión (o masc.)



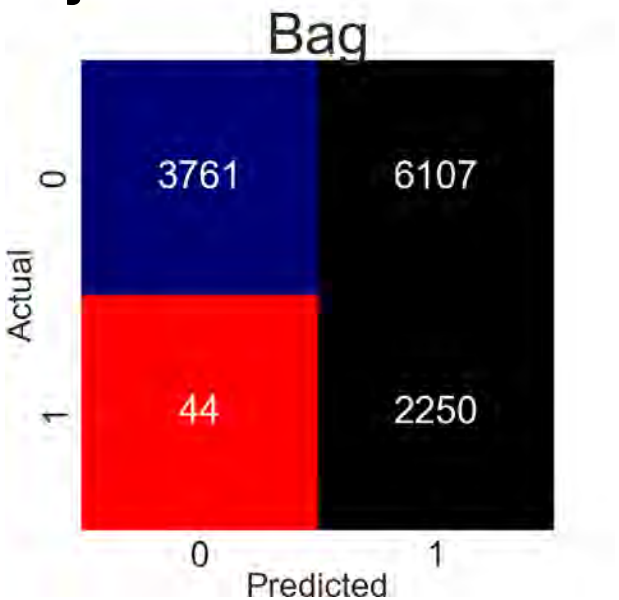
Errores

Annotated female
Predicted as male

Annotated male
Predicted as female

Visión y lenguaje

[Batallita]
Matriz de confusión (1 presente)



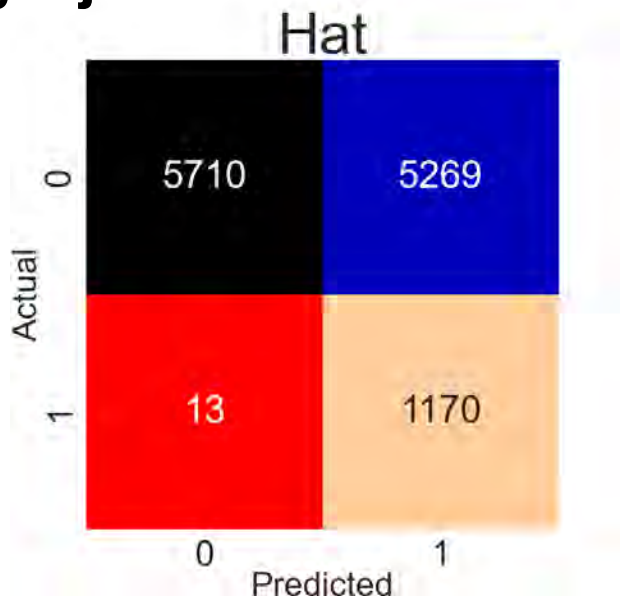
Errores

Annotated no bag
Predicted bag

Annotated bag
Predicted no bag

Visión y lenguaje

[Batallita]
Matriz de confusión (1 presente)



Errores

Annotated no hat
Predicted hat

Annotated hat
Predicted no hat

Beanies?

Visión y lenguaje

[Batallita]

Resultados con *baseline* conjunto privado

$$Acc. = \frac{\sum_1^K (p_i == g_i)}{K}$$

Approach	Upper color	Lower color	Gender	Bag	Hat	Mean
Baseline	0.628	0.724	0.721	0.641	0.832	0.709

Visión y lenguaje

[Batallita]

Resultados con *baseline* conjunto privado

$$Acc. = \frac{\sum_1^K (p_i == g_i)}{K}$$

Approach	Upper color	Lower color	Gender	Bag	Hat	Mean
Baseline	0.628	0.724	0.721	0.641	0.832	0.709
ROC-ULPGC	0.9207	0.9081	0.9272	0.9215	0.9279	0.9211

$Rank$	$Team$	A_u	A_l	A_g	A_b	A_h	mA
1	IROC-ULPGC	0.9207	0.9081	0.9272	0.9215	0.9279	0.9211
2	HUSTNB	0.7284	0.7478	0.8448	0.6308	0.8678	0.7639
3	SPARKY	0.7387	0.9041	0.8833	0.6635	0.5772	0.7533
4	CODY	0.5970	0.7431	0.8478	0.4463	0.5438	0.6356
5	AWESOMEPAR	0.1817	0.1377	0.9358	0.7712	0.7559	0.5565

Visión y lenguaje

[Batallita]

OPT, PnP y FLANT5XL (*accuracy/precision/recall/F1-score*)

CAIP 23 task	OPT	PNP	FLANT5XL
Upper color	**0.804**/**0.804**/**0.804**/**0.801**	0.091/0.298/0.092/0.116	0.602/0.820/0.602/0.620
Lower color	**0.842**/**0.844**/**0.842**/**0.839**	0.093/0.468/0.093/0.145	0.493/0.587/0.493/0.522
Gender	0.909/0.915/0.754/0.827	0.287/0.287/1.000/0.446	**0.958**/**0.959**/**0.891**/**0.924**
Bag	0.494/0.269/**0.980**/0.422	0.349/0.218/0.946/0.354	**0.885**/**0.748**/0.595/**0.661**
Hat	0.566/0.181/**0.987**/0.307	0.128/0.099/0.992/0.181	**0.923**/**0.688**/0.374/**0.484**

M. Castrillón et al. *Visual Question Answering Models for Zero-Shot Pedestrian Attribute Recognition: A Comparative Study.* SN Computer Science, 2024

Visión y lenguaje

[Batallita]

OPT, PnP y FLANT5XL (*accuracy/precision/recall/F1-score*)

CAIP 23 task	OPT	PNP	FLANT5XL
Upper color	**0.804**/**0.804**/**0.804**/**0.801**	0.091/0.298/0.092/0.116	0.602/0.820/0.602/0.620
Lower color	**0.842**/**0.844**/**0.842**/**0.839**	0.093/0.468/0.093/0.145	0.493/0.587/0.493/0.522
Gender	0.909/0.915/0.754/0.827	0.287/0.287/1.000/0.446	**0.958**/**0.959**/**0.891**/**0.924**
Bag	0.494/0.269/**0.980**/0.422	0.349/0.218/0.946/0.354	**0.885**/**0.748**/0.595/**0.661**
Hat	0.566/0.181/**0.987**/0.307	0.128/0.099/0.992/0.181	**0.923**/**0.688**/0.374/**0.484**

Procesamiento (msegs.)

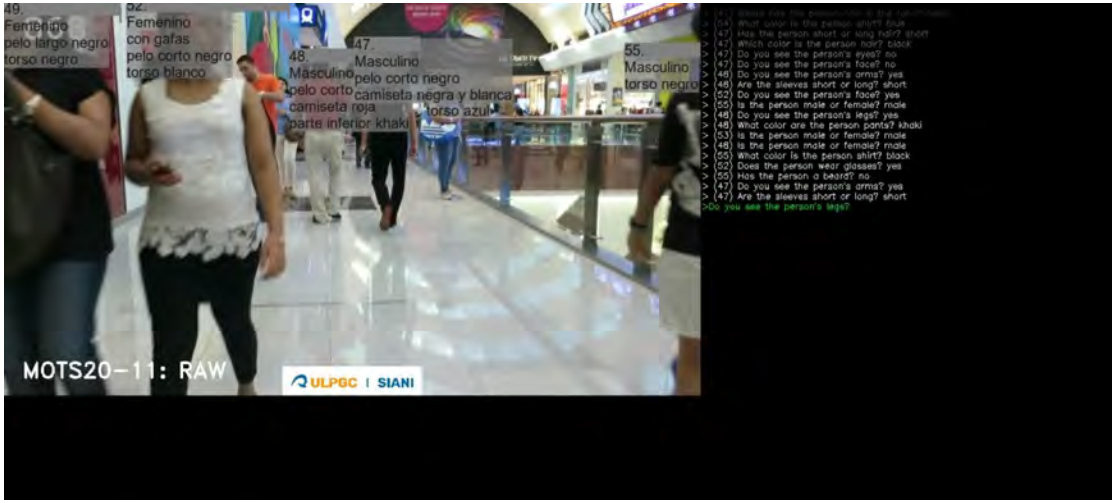
A40

Question	OPT	PNP	FLANT5XL
Does the person wear a jacket?	0.0149	0.0993	0.0518
What color is the person's jacket?	0.0131	0.0984	0.0265
What color is the person's shirt?	0.0130	0.0982	0.0230
What color is the person's trousers?	0.0130	0.0978	0.0270
Is the person male or female?	0.0156	0.0989	0.0317
Does the person wear a bag?	0.0127	0.0978	0.0539
Does the person wear a hat?	0.0150	0.0982	0.0543
Does the person wear a cap?	0.0148	0.0986	0.0545

M. Castrillón et al. *Visual Question Answering Models for Zero-Shot Pedestrian Attribute Recognition: A Comparative Study.* SN Computer Science, 2024

Visión y lenguaje

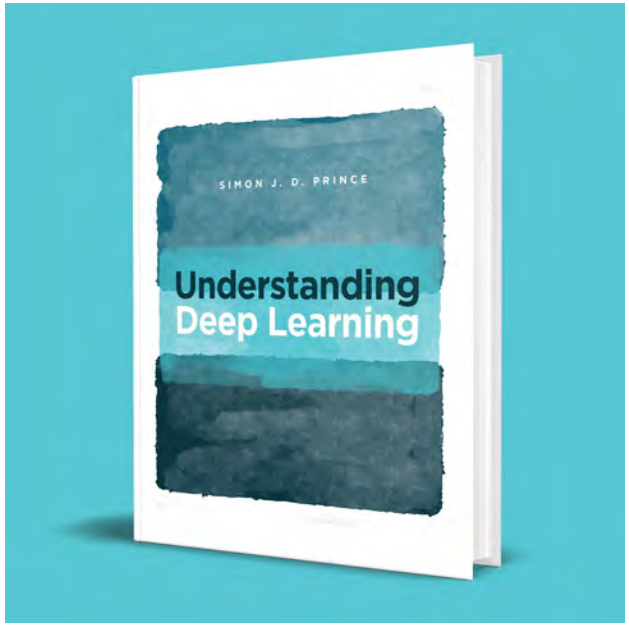
[Batallita]



Visión y lenguaje

[Batallita]
¿Cómo va BLIP-2 OPT con ...?

- How many cars are in the image? **1**
- How many wheels are visible? **2**
- Is any light on? **no**
- Where is the car from? **germany**
- Which color is the car? **red**
- What shade of red? bright **red**
- Is the license plate visible? **yes**
- What is written in the license plate? **gj58ltv**
- Is the car front visible? **yes**
- Which side of the car is visible? **Right**

Bibliografía

- Mikael Nilsson. Face Detection
- Simon J. D. Prince. Understanding Deep Learning, 2024
- R. Szeliski. Computer Vision: Algorithms and Applications, 2nd ed. 2022
- A. Torralba et al. Foundations of Computer Vision, 2024