

Sistema de Detección de Líneas de Pista y Seguimiento Automático de Jugadores y Pelota en Partidos de Tenis en Tiempo Real

Marino Fernández Pérez

marinofepe@correo.ugr.es

Francesc Oliver Catany

francescoliver@correo.ugr.es

Pau Bover Femenias

pauboover@correo.ugr.es

Gabriele Ruggeri

gabricross37@correo.ugr.es

Universidad de Granada (UGR) - Visión por Computador

Abstract

En este trabajo se presenta un sistema para el análisis automático de un partido de tenis a partir de vídeo broadcast. El sistema es capaz de identificar la pista de juego, sus líneas y keypoints relevantes, así como detectar y seguir a los jugadores y la pelota a lo largo de la secuencia de vídeo. A partir de esta información, se estiman métricas cinemáticas como la velocidad y la distancia recorrida por los jugadores durante el partido.

La solución propuesta integra múltiples modelos de aprendizaje profundo junto con técnicas clásicas de visión por computador, organizados en un pipeline modular y extensible. El sistema ha sido evaluado en vídeos reales de diferentes competiciones y tipos de pista, mostrando un funcionamiento robusto en escenarios variados. Aunque el procesamiento se realiza actualmente en un régimen de casi tiempo real, se discute su viabilidad para aplicaciones en tiempo real mediante optimizaciones adicionales. Los resultados obtenidos demuestran el potencial del enfoque propuesto como herramienta de análisis automático en el ámbito del tenis profesional.

1. Introducción

En el ámbito del tenis profesional, el análisis detallado del rendimiento tanto de jugadores propios como de rivales supone una tarea compleja y costosa para entrenadores y analistas. Gran parte de este análisis se realiza de forma manual o semiautomática, lo que dificulta la obtención de información objetiva y detallada a partir de grandes volúmenes de vídeo.

La resolución de este problema resulta relevante debido al creciente interés en el uso de herramientas basadas

en inteligencia artificial para el análisis avanzado del rendimiento deportivo. La información recopilada por sistemas automáticos de detección y seguimiento permite desarrollar métricas complejas y estimaciones sobre el comportamiento de los jugadores durante el partido. Este tipo de análisis puede proporcionar una ventaja competitiva en el estudio de rivales, facilitando la identificación de patrones de juego y tendencias estratégicas. Además, este tipo de tecnología presenta un alto potencial de aplicación en retransmisiones televisivas, donde puede emplearse para ofrecer información visual y estadística de interés que enriquezca la experiencia de los espectadores.

El objetivo de este trabajo es desarrollar un sistema que permita detectar de manera automática información relevante a partir de partidos de tenis, de forma que cualquier usuario pueda utilizarlo para analizar vídeos y construir métricas y análisis avanzados basados en dicha información.

1.1. Motivación personal

Elegimos este tema como proyecto porque nos pareció un campo especialmente amplio y estimulante, en el que era posible profundizar en muchos de los conceptos que más nos habían interesado a lo largo de la asignatura. El análisis de vídeo deportivo combina técnicas de aprendizaje profundo, procesamiento clásico de imágenes y geometría, lo que nos permitió aplicar de forma práctica conocimientos diversos que iremos detallando más adelante. Además, la complejidad real del problema y la necesidad de tomar decisiones de diseño fundamentadas lo convirtieron en un contexto idóneo para consolidar y ampliar lo aprendido durante el curso.

2. Contexto

El análisis automático de partidos de tenis a partir de secuencias de vídeo constituye un problema relevante dentro del ámbito de la visión por computador, requiere la detección y el seguimiento de elementos dinámicos, como jugadores y pelota, así como la localización de estructuras geométricas estáticas, en particular la pista. Para abordar estos retos de forma robusta, resulta fundamental combinar técnicas de aprendizaje profundo con métodos clásicos de procesamiento de imagen.

En los últimos años, los modelos basados en aprendizaje profundo han mostrado un alto rendimiento en tareas de detección y seguimiento de objetos en tiempo real. Arquitecturas como YOLO se utilizan habitualmente para la detección de jugadores, mientras que redes especializadas como TrackNet han sido diseñadas para el seguimiento de objetos pequeños en movimiento, como la pelota. Asimismo, modelos de aprendizaje profundo también se emplean para la detección de puntos clave relevantes de la pista. El uso de modelos ligeros permite además realizar inferencia eficiente, facilitando su integración en sistemas de tiempo real.

Por otro lado, la detección de las líneas de la pista presenta características distintas, ya que se trata de estructuras geométricas bien definidas y con una disposición regular. En este contexto, los métodos clásicos de procesamiento de imagen, como la Transformada de Hough, continúan siendo una alternativa eficaz, especialmente cuando se combinan con técnicas de preprocesado que reducen el ruido y mejoran la calidad geométrica.

Finalmente, la relación entre la imagen capturada por la cámara y el plano real de la pista puede modelarse mediante técnicas de geometría proyectiva. El uso de homografías permite establecer una correspondencia entre ambos planos, posibilitando la superposición de modelos de referencia y la interpretación espacial de las posiciones detectadas.

3. Trabajos Previos

El análisis automático de partidos de tenis ha despertado un creciente interés en los últimos años, impulsado por los avances en visión por computador y aprendizaje profundo. De forma general, los trabajos existentes pueden agruparse en tres líneas principales: el seguimiento de la pelota, la detección y seguimiento de jugadores, y la detección de la pista y sus líneas.

Para el seguimiento de la pelota, uno de los enfoques más influyentes es *TrackNet* [1], una red neuronal profunda que formula el problema como una tarea de segmentación mediante mapas de probabilidad. Este enfoque ha demostrado una elevada precisión en el seguimiento de

objetos pequeños y de alta velocidad, como la pelota de tenis, incluso en vídeos broadcast de resolución y tasa de frames moderadas, consolidándose como una referencia en este ámbito.

En cuanto a la detección de jugadores, es habitual el uso de detectores de propósito general basados en aprendizaje profundo, como YOLO [3] o variantes de *Faster R-CNN*, que ofrecen un buen compromiso entre precisión y eficiencia en escenarios reales. Por otro lado, la detección de la pista suele abordarse mediante enfoques híbridos que combinan aprendizaje profundo y técnicas clásicas de procesamiento de imagen. Trabajos como ML6 [2] y diversas implementaciones abiertas [4, 5] emplean redes neuronales para una detección inicial robusta, complementada posteriormente con razonamiento geométrico para preservar la coherencia estructural de la pista. En conjunto, estos trabajos evidencian que la combinación de múltiples técnicas resulta clave para obtener sistemas robustos y precisos en el análisis automático de tenis.

4. Metodología

4.1. Detección de la pista de tenis

La detección de la pista de tenis se aborda como un problema de localización geométrica de estructuras lineales estáticas dentro de una escena de vídeo.

4.1.1. Enfoque de IA simbólica (Hough Lines)

Antes de adoptar el enfoque híbrido, se exploró una solución basada exclusivamente en técnicas geométricas clásicas. En concreto, el procedimiento consistía en convertir la imagen a escala de grises, aplicar un umbral fijo para resaltar las líneas claras de la pista, suavizar el resultado mediante un filtrado gaussiano y extraer bordes con el detector de Canny. Sobre esta imagen de bordes se aplicaba posteriormente la Transformada de Hough probabilística con el objetivo de detectar segmentos rectos correspondientes a las líneas de la pista.

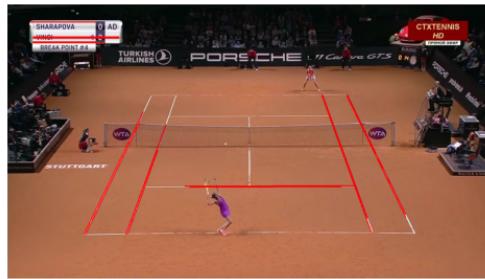


Figure 1. Primer enfoque usando técnicas clásicas

Este enfoque presenta diversas limitaciones, debido a que dependen en gran medida de supuestos idealizados,

como una buena separación entre líneas y fondo, iluminación homogénea y ausencia de elementos distractores. En la práctica, las variaciones de iluminación, las sombras, los reflejos, etc provocan que las líneas no se destaque en forma uniforme frente al fondo, dificultando su segmentación mediante umbrales o detectores de bordes.

Además, la perspectiva de la cámara introduce cambios significativos en el grosor y la continuidad aparente de las líneas, especialmente en zonas lejanas, donde estas se vuelven muy finas. En este tipo de situaciones, los detectores basados en bordes y la Transformada de Hough tienden a producir detecciones fragmentadas o incompletas. A esto se suma la interferencia de otros elementos de la escena, como jugadores, red, marcadores gráficos, que generan bordes que no representan líneas reales.

Como resultado, en imágenes complejas, la detección puramente geométrica falla al no ser capaz de discriminar de forma fiable las líneas de la pista. Estas limitaciones ponen de manifiesto la necesidad de complementar los métodos clásicos con técnicas basadas en aprendizaje profundo, que aporten una mayor robustez frente a variaciones visuales y permitan guiar o refinar posteriormente el análisis geométrico.

4.1.2. Enfoque con CNN, para detección de keypoints

Posteriormente, se optó por explorar un enfoque alternativo basado en redes neuronales convolucionales, propuesto en ML6 [2] , en el que una CNN se emplea para estimar directamente la posición de los keypoints de la pista. Este enfoque permite modelar de forma implícita variaciones complejas de iluminación, perspectiva y occlusiones, superando algunas de las limitaciones de los métodos puramente geométricos.

En el artículo mencionado se experimenta con distintas arquitecturas de redes neuronales y configuraciones de entrenamiento, concluyéndose que el modelo que ofrece un mejor equilibrio entre precisión y eficiencia es **MobileNetV3-Small**. A partir de esta conclusión, se decidió replicar dicho planteamiento como punto de partida, con el objetivo de validar sus resultados y, posteriormente, explorar posibles mejoras.

El entrenamiento para este problema de regresión se realiza mediante una estrategia en dos fases, comenzando con un ajuste de la cabecera del modelo y seguido de un *fine-tuning* completo de la red.

Si bien los resultados obtenidos mediante *fine-tuning* eran satisfactorios y el modelo era capaz de capturar correctamente los patrones de la pista, se observó que los keypoints estimados no quedaban posicionados con suficiente precisión sobre las intersecciones reales de las líneas 2 . Esta falta de alineación geométrica provocaba que, al

NETWORK PARAMETERS OF THE MOBILENETV3-SMALL ARCHITECTURE USED FOR COURT KEYPOINT DETECTION

Layer	FS	Depth	Padding	Stride	Activation
Stem Conv	3×3	16	1	2	HSW+BN
IR Block 1	3×3	16	1	2	ReLU / HSW
IR Blocks 2–3	3×3	24	1	2 / 1	ReLU
IR Blocks 4–6	3×3	40	1	2 / 1	HSW + SE
IR Blocks 7–8	3×3	48	1	1	HSW + SE
IR Blocks 9–11	3×3	96	1	2 / 1	HSW + SE
Conv Final	1×1	576	0	1	HSW + BN
Global Pool					Adaptive Average Pooling
FC1	–	1024	–	–	HSW
Dropout	–	–	–	–	Dropout
Output	–	28	–	–	Linear

Nota: HSW refiere a la función de activación Hard-Swish.

aplicar etapas posteriores basadas en dichos keypoints, los resultados no fueran plenamente fiables ni consistentes. Por este motivo, se consideró necesario introducir una etapa adicional de postprocesado, orientada a corregir y refinar la posición de los keypoints, con el fin de mejorar su coherencia geométrica.

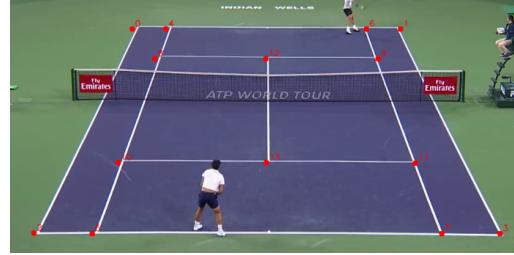


Figure 2. Keypoints detectados sin refinamiento.

4.1.3. Enfoque con CNN y postprocesado

El postprocesado comienza con la extracción de regiones de interés (crops) centradas en keypoints detectados previamente. Estos keypoints proporcionan una aproximación inicial a la localización de las líneas, usamos la red como estimador de posiciones. Cada crop se procesa de forma independiente con el objetivo de refinar la posición del punto mediante la geometría local de las líneas.

En primer lugar, se aplica un preprocessado orientado a la detección de líneas que incluye la reducción del espacio de color, la binarización y una dilatación adaptativa para reforzar la continuidad de las líneas. Posteriormente, se utiliza el algoritmo de afinado de Zhang–Suen para obtener un esqueleto de un solo píxel de grosor, lo que mejora de forma significativa la estabilidad y precisión de la Transformada de Hough al eliminar ambigüedades asociadas a líneas anchas o bordes difusos.

Sobre la imagen afinada se aplica la Transformada de Hough probabilística para detectar segmentos de línea recta, garantizando que las detecciones correspondan a estruc-

turas reales de la pista y no a ruido. Finalmente, las líneas obtenidas se filtran y agrupan según su orientación y proximidad, fusionando aquellas que representan la misma estructura física.

Una vez obtenidas las líneas coherentes dentro de cada crop, se calculan sus intersecciones. Estas intersecciones representan puntos geométricamente estables, como cruces o esquinas de las líneas de la pista, y constituyen una referencia mucho más fiable que los keypoints iniciales. Dado que en un mismo crop pueden existir múltiples intersecciones —especialmente en presencia de la red, que es detectada como una línea válida—, se selecciona como punto refinado aquella intersección más cercana al keypoint original. Este criterio de proximidad ha demostrado ser robusto en la mayoría de los casos, incluso en escenarios complejos. Finalmente, los puntos finales, inicialmente expresados

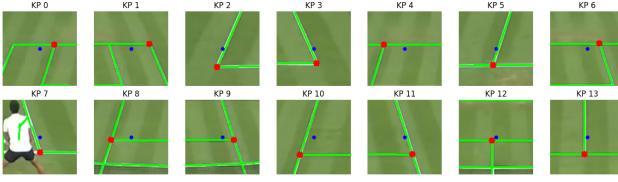


Figure 3. Keypoints CNN: azul - Keypoints Refinados: rojo

en el sistema de coordenadas local del crop, se transforman al sistema de coordenadas global de la imagen. Con este proceso obtenemos puntos coherentes sobre las que aplicar homografías y cálculos posteriores.

4.2. Homografía de las líneas

Una vez obtenidos los keypoints correctos, el siguiente paso consiste en proyectar las líneas de la pista sobre la imagen original. Para ello se emplea una homografía, que permite establecer una correspondencia geométrica precisa entre los keypoints detectados en la imagen y una pista de referencia definida en un sistema de coordenadas conocido. Esta transformación modela la deformación proyectiva introducida por la cámara y hace posible trasladar una plantilla ideal de la pista, construida a partir de las dimensiones oficiales del tenis y escalada a píxeles, sobre la escena real. El cálculo de la homografía se realiza mediante **RANSAC**, lo que permite estimar una transformación robusta incluso en presencia de keypoints erróneos. Gracias a este enfoque, se obtiene una superposición coherente de la pista que no solo facilita la visualización, sino que constituye la base para tareas posteriores como la localización precisa de los jugadores sobre la pista o el cálculo de distancias en escala real.

4.3. Detección y seguimiento de jugadores

Una vez obtenida una estimación precisa y geométricamente consistente de la pista de tenis, el siguiente paso del sistema consiste en la detección y el

seguimiento de los jugadores a lo largo de la secuencia de vídeo. A diferencia de la detección de la pista, que se basa en estructuras estáticas, esta etapa aborda el análisis de elementos dinámicos, lo que introduce desafíos adicionales relacionados con occlusiones, cambios de postura y variaciones de escala debidas a la perspectiva de la cámara.

La detección inicial se realiza mediante un detector basado en aprendizaje profundo, concretamente el modelo **YOLO**, entrenado para la detección de objetos en imágenes en tiempo real, y en este caso, de personas. Para cada frame del vídeo, el detector proporciona un conjunto de *bounding boxes* correspondientes a todas las personas visibles, incluyendo tanto a los jugadores como a otros individuos presentes en la pista, como jueces de línea o recogepelotas, lo que se convierte en el principal problema.

Por tanto, la dificultad no reside únicamente en la detección de personas, sino en la selección consistente de los dos tenistas relevantes y en el mantenimiento de su identidad (jugador superior e inferior) a lo largo del clip. Para ello, se exploraron de forma incremental distintos enfoques, analizando sus limitaciones y refinándolos progresivamente hasta alcanzar la estrategia final empleada en el sistema.

4.3.1. Enfoque con filtrado a partir de ROI

Se introduce un filtrado espacial basado en regiones de interés (*Region of Interest*, ROI). Dado que en las retransmisiones de tenis la cámara suele permanecer aproximadamente fija y la pista ocupa una zona bien definida dentro de la imagen, se asume que los jugadores relevantes aparecen mayoritariamente dentro de dicha región.

Con ello, se define una ROI fija que abarca el área principal de la pista y respeta su forma trapezoidal, descartando automáticamente todas aquellas detecciones de personas situadas fuera de ella.

Sin embargo, presenta varias limitaciones. La definición de una ROI fija resulta altamente dependiente del encuadre concreto del vídeo, lo que dificulta su generalización a distintas retransmisiones o a variaciones en el ángulo de cámara. Además, en situaciones como cambios de plano, frames en los que la posición de los tenistas se desvía del patrón habitual o encuadres más abiertos, los jugadores pueden quedar parcial o totalmente fuera de la región definida, provocando pérdidas en la detección.

Asimismo, incluso dentro de la ROI continúan apareciendo detecciones no deseadas, como recogepelotas que acceden temporalmente a la pista o jueces visibles en determinados planos. Esto pone de manifiesto que el filtrado puramente espacial no resulta suficiente para garantizar una selección robusta y consistente a variaciones visuales.



Figure 4. Limitaciones con ROI

4.3.2. Enfoque basado en keypoints reales de la pista

Como alternativa, el uso de la información geométrica proporcionada por los puntos de interés reales de la pista de tenis permite representar posiciones estructurales relevantes del campo y definir referencias directamente ligadas a la geometría del escenario de juego, independientemente de cómo esté situada la cámara.



Figure 5. Selección de jugadores usando la información geométrica de la pista

Ahora, para cada persona detectada se calcula un punto representativo asociado a su posición en el plano de la imagen, definido a partir del centro de su *bounding box*, y se mide la distancia euclídea entre dicho punto y el conjunto de keypoints de la pista, considerando la distancia mínima como un indicador de cercanía al área de juego. Intuitivamente, se asume que los jugadores relevantes serán aquellos dos individuos cuya distancia mínima a la estructura de la pista sea menor.

No obstante, el comportamiento de esta estrategia se ve condicionado por la distribución espacial de los keypoints de la pista. Al concentrarse principalmente en las líneas laterales, líneas de fondo e intersecciones, pueden darse situaciones en las que personas situadas cerca de estas zonas, como recogepelotas próximos a la red, resulten geométricamente más cercanas a un keypoint que un jugador correctamente posicionado en el interior del campo.

Este efecto se ve reforzado por la perspectiva de la cámara, especialmente en la mitad superior de la pista, donde la compresión visual reduce las distancias aparentes entre elementos externos y la zona de juego. Además, este criterio no garantiza la selección de un jugador por cada

mitad del campo, pudiendo asignar ambos candidatos a una misma región en determinados frames.

4.3.3. Enfoque basado en la definición de keypoints virtuales

Para mejorar la selección de jugadores y evitar las limitaciones asociadas al uso directo de los keypoints reales de la pista, se introduce un enfoque basado en la definición de *keypoints virtuales*. Estos puntos se sitúan en regiones representativas de las zonas donde los jugadores suelen posicionarse durante el juego, concretamente cerca de las líneas de fondo de cada mitad de la pista.

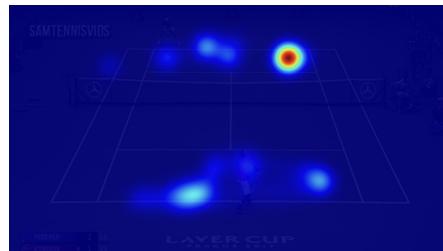


Figure 6. Mapa de calor del patrón habitual de movimiento

A partir de estos keypoints virtuales, la asignación de jugadores se formula como un problema de minimización de coste en función de su distancia a estos, obviando los puntos de interés reales y garantizando explícitamente la selección de un jugador por cada mitad del campo.



Figure 7. Selección de jugadores con puntos de interés virtuales

Al utilizar referencias geométricas más estables y semánticamente significativas se observan mejoras claras. Sin embargo, los errores persisten principalmente en la mitad superior de la pista, incluso cuando los keypoints virtuales están correctamente definidos. La causa de estas limitaciones radica en que, una vez más, el razonamiento geométrico se realiza en el plano imagen, donde la distancia aparente no refleja la distancia real sobre la pista debido a la perspectiva de la cámara.

4.3.4. Enfoque basado en homografía inversa

Trasladar la asignación de jugadores al plano real de la pista mediante el uso de la homografía inversa permite proyectar puntos desde la imagen al sistema de referencia de la pista,

donde las distancias, ahora sí, representan desplazamientos reales y no están afectadas por la perspectiva de la cámara.

El procedimiento de selección es idéntico al del enfoque anterior, aunque en lugar de utilizar el centro geométrico de la *bounding box*, se emplea un punto situado en el centro de su borde inferior, que aproxima la posición real de los pies del jugador sobre la pista, garantizando la validez de la proyección mediante la homografía inversa al ser el punto más representativo con respecto al plano pista.

Al operar en un espacio geométricamente coherente, se elimina en gran medida el impacto distorsionador de la perspectiva, lo que resulta especialmente beneficioso para la mitad superior de la pista, donde los errores eran más frecuentes. Con este proceso la selección de ambos jugadores es más estable y consistente, validando que el uso combinado de keypoints virtuales y homografía inversa constituye la estrategia más robusta dentro del sistema propuesto.

4.4. Representación de jugadores en un minimapa de la pista

Una vez establecida la correspondencia geométrica entre la imagen original y el plano de referencia de la pista mediante homografía, se introduce una representación auxiliar en forma de minimapa, este se construye a partir de la plantilla de la pista en vista cenital, definida en un sistema de coordenadas conocido y se amplía mediante un margen extendido alrededor de la pista, con el fin de representar de forma consistente posiciones situadas fuera del rectángulo de juego.

Las posiciones de los jugadores, se proyectan al plano de la pista mediante la homografía inversa de la calculada anteriormente. Para ello, las coordenadas en el sistema de referencia de la pista se transforman al sistema del minimapa teniendo en cuenta el factor de escala y el desplazamiento asociado al margen extendido. Los puntos se dibujan sobre la capa final del minimapa, evitando así que se vean afectados por el proceso de transparencia y asegurando una visualización clara y estable.

Un aspecto fundamental del procedimiento es la selección del punto representativo de cada jugador. En este trabajo se emplea el punto medio del borde inferior de la bounding box, que aproxima la posición de los pies del jugador sobre la superficie de la pista. Esta elección resulta crucial desde el punto de vista geométrico, ya que dicho punto pertenece al plano físico sobre el que se ha estimado la homografía. El uso de puntos elevados, como el centro del cuerpo, introduce errores de proyección ya que no se encuentra sobre el plano donde se ha calculado la homografía. Esta visualización proporciona una interpretación espacial clara del movimiento de los jugadores y constituye una base sólida para el cálculo posterior de métricas cinemáticas.

4.5. Seguimiento de la Pelota

El seguimiento de la pelota de tenis se aborda mediante TrackNet [1]. A diferencia de los detectores estáticos convencionales, TrackNet explota de forma explícita la información temporal, lo que resulta esencial en un escenario como el tenis, donde la pelota ocupa muy pocos píxeles y presenta movimientos rápidos. El problema se formula como una tarea de segmentación mediante mapas de calor (*heatmaps*), en lugar de una regresión directa de coordenadas (x, y). Para cada secuencia de entrada, el modelo produce un mapa de probabilidad en el que la máxima activación indica la posición estimada de la pelota.

NETWORK PARAMETERS OF THE PROPOSED TRACKNET ARCHITECTURE

Layer	Filter Size	Depth	Padding	Stride	Activation
Conv1-1	3×3	64	1	1	ReLU+BN
Conv1-2	3×3	64	1	1	ReLU+BN
Pool1	2×2 max pooling and <i>Stride</i> =2				
Conv2-1	3×3	128	1	1	ReLU+BN
Conv2-2	3×3	128	1	1	ReLU+BN
Pool2	2×2 max pooling and <i>Stride</i> =2				
Conv3-1	3×3	256	1	1	ReLU+BN
Conv3-2	3×3	256	1	1	ReLU+BN
Pool3	2×2 max pooling and <i>Stride</i> =2				
Bneck-1	3×3	512	1	1	ReLU+BN
Bneck-2	3×3	512	1	1	ReLU+BN
UpS1	bilinear upsampling ($\times 2$)				
Conv5-1	3×3	256	1	1	ReLU+BN
Conv5-2	3×3	256	1	1	ReLU+BN
UpS2	bilinear upsampling ($\times 2$)				
Conv6-1	3×3	128	1	1	ReLU+BN
Conv6-2	3×3	128	1	1	ReLU+BN
UpS3	bilinear upsampling ($\times 2$)				
Conv7-1	3×3	64	1	1	ReLU+BN
Conv7-2	3×3	64	1	1	ReLU+BN
Output	1×1	1	0	1	Linear

Nota: La arquitectura sigue un diseño de tipo U-Net. Las skip connections se aplican entre el codificador y el decodificador mediante concatenación a nivel de canales.

Aunque TrackNet proporciona una localización precisa de la pelota en la mayoría de los casos, se observaron activaciones espurias y inconsistencias temporales en escenarios complejos; para corregir estas limitaciones, se aplica una etapa ligera de postprocesado sobre los mapas de calor predichos, que incluye la alineación espacial de las detecciones, el filtrado lógico y la interpolación temporal de posiciones ausentes, así como el suavizado de la trayectoria, mejorando significativamente la estabilidad temporal y reduciendo los falsos positivos.

5. Experimentos

I. DATASET

Para el entrenamiento y la evaluación del sistema se han utilizado dos conjuntos de datos públicos orientados al análisis automático de partidos de tenis a partir de vídeo. En primer lugar, se emplea el *TrackNet Tennis Dataset*¹, compuesto por secuencias de frames con resolución fija de 640×360 píxeles y anotaciones frame a frame de la posición de la pelota (x, y).

Asimismo, se utiliza el dataset asociado al proyecto *TennisCourtDetector*², formado por 8 841 imágenes de resolución 1280×720 que cubren distintos tipos de superficie e incluyen la anotación de 14 puntos clave por imagen que describen la geometría de la pista.

La combinación de ambos datasets permite evaluar el sistema en escenarios realistas, integrando información dinámica (pelota y jugadores) y estática (pista) dentro de un único pipeline experimental.

II. PLANTEAMIENTO EXPERIMENTAL

TrackNet Tennis Dataset se emplea tanto para el entrenamiento del modelo de detección de la pelota como para la evaluación del resto de módulos del sistema sobre clips completos de partidos reales.

La partición de los datos se realiza a nivel de partido (*game*), asignando cada juego completo al conjunto de entrenamiento o validación garantizando que no existan frames del mismo juego en ambos conjuntos.

TRAINING PARAMETERS FOR TRACKNET ARCHITECTURE

Parameter	Value
Input resolution	640×360
Batch size	32
Optimizer	Adam
Initial LR	1×10^{-4}
Epochs	50
Loss function	Focal loss (heatmaps)

Durante el entrenamiento se aplica *data augmentation* ligero sobre las secuencias de entrada, incluyendo variaciones aleatorias de brillo y contraste, así como la adición de ruido gaussiano, con el objetivo de mejorar la robustez del modelo frente a cambios de iluminación y calidad de imagen.

Para mitigar el fuerte desbalance entre el fondo y la región correspondiente a la pelota en los mapas de calor, se utiliza *Focal Loss*, formulada a partir de *Binary Cross-Entropy*, BCE:

¹<https://www.kaggle.com/datasets/sofuskonglevoll/tracknet-tennis>

²<https://github.com/yastrebksv/TennisCourtDetector/tree/main>

$$\mathcal{L}_{\text{focal}} = \alpha (1 - p_t)^\gamma \mathcal{L}_{\text{BCE}}, \quad \text{con } p_t = \exp(-\mathcal{L}_{\text{BCE}})$$

Esta función penaliza con mayor peso los errores en las regiones difíciles y reduce la contribución de los píxeles correctamente clasificados.

Para la detección de keypoints de la pista, *TennisCourt-Detector* se divide en un 75% para entrenamiento y un 25% para validación, siguiendo la partición propuesta en el dataset original.

TRAINING PARAMETERS FOR COURT KEYPOINT DETECTION

Parameter	Value
Input resolution	480×480
Batch size	16
Optimizer	Adam
Initial LR	1×10^{-3}
Epochs (head only)	5
Epochs (full model)	20
Loss function	L1 Loss (MAE)

Para evaluarlo el entrenamiento, se mide el error promedio absoluto entre las posiciones predichas por el modelo y las posiciones reales anotadas en el dataset (*MAE*):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$$

III. EXPERIMENTOS Y RESULTADOS

Para TrackNet, el comportamiento de la pérdida frente al número de épocas muestra una convergencia estable a lo largo del entrenamiento. El uso de secuencias de tres frames consecutivos como entrada permite explotar la información temporal, mejorando la robustez del modelo frente a trayectorias rápidas y occlusiones parciales.

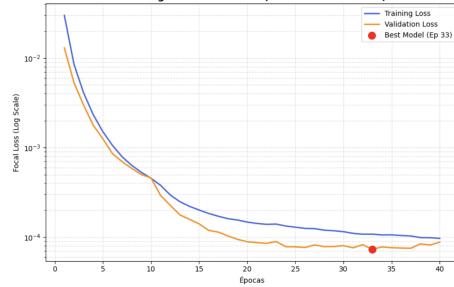


Figure 8. Curva de aprendizaje para TrackNet

Como se observa en la Fig. 8, la pérdida de validación se mantiene por debajo de la pérdida de entrenamiento debido al uso de *data augmentation* durante la fase de entrenamiento, cuyo conjunto presenta una mayor dificultad que el conjunto de validación, que se evalúa sin transforma-

ciones.

El entrenamiento se prolonga hasta aproximadamente la época 40, momento en el que se activa el criterio de *early stopping* al no observarse mejoras adicionales en la pérdida de validación. Esta estrategia permite prevenir el sobreajuste y seleccionar un modelo con buena capacidad de generalización.

En el caso de la detección de keypoints de la pista, se evalúa el impacto de distintas estrategias de entrenamiento sobre el modelo.

TRAINING RESULTS FOR COURT KEYPOINT DETECTION

Training Strategy	Final Loss (MAE)
From scratch	12.4700
Head only	10.5181
Full fine-tuning	7.6267

Los resultados indican que, aun partiendo de pesos preentrenados, la estrategia de entrenamiento influye de forma significativa en el rendimiento final. El ajuste exclusivo de la cabecera produce una mejora limitada, mientras que el entrenamiento completo desde el inicio resulta menos estable. En contraste, el *fine-tuning* progresivo, comenzando por la cabecera y extendiéndose posteriormente a toda la red, alcanza el menor *MAE*.

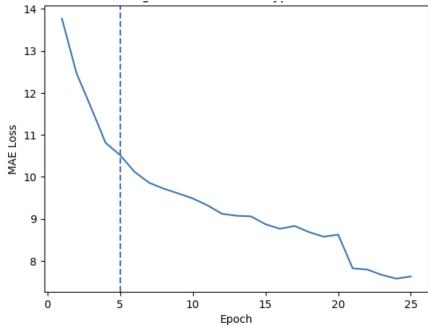


Figure 9. Curva de aprendizaje para fine-tuning completo

En adición a la evaluación de los modelos individuales, se ha definido una métrica orientada a cuantificar la distancia acumulada recorrida de los jugadores a lo largo del clip. Dado que el sistema opera sobre imágenes, la distancia se calcula inicialmente en píxeles y se transforma posteriormente a metros mediante una relación de escala aproximada basada en la altura estimada del jugador (1.80). Para garantizar la estabilidad de la señal, se emplean posiciones interpoladas en aquellos frames donde la detección no es válida, evitando discontinuidades en el cálculo.

6. Conclusiones

En este trabajo se ha presentado un sistema para el análisis automático de partidos de tenis a partir de vídeo broad-

cast, basado en la integración de modelos de aprendizaje profundo con técnicas clásicas de visión por computador y geometría proyectiva. El sistema es capaz de detectar la pista y sus keypoints geométricos, así como de identificar y seguir a los jugadores y la pelota a lo largo de la secuencia de vídeo de forma robusta. Los resultados experimentales muestran que la combinación de redes neuronales convolucionales con etapas de postprocesado geométrico permite mejorar significativamente la coherencia espacial y la estabilidad de las detecciones frente a enfoques puramente basados en aprendizaje profundo. En particular, el uso de homografías y referencias geométricas de la pista resulta clave para desacoplar el razonamiento espacial de los efectos de la perspectiva de la cámara, permitiendo una asignación de jugadores más fiable y consistente.

Asimismo, el uso de modelos ligeros demuestra que es posible obtener un rendimiento elevado en escenarios reales sin necesidad de vídeo de alta tasa de frames ni resoluciones extremas, reduciendo el coste computacional y facilitando su aplicación práctica. Sin embargo, a pesar de nuestros esfuerzos, el pipeline completo no consigue el rendimiento necesario para poderse aplicar en tiempo real, aunque con optimizaciones y una mayor potencia de cómputo podría reducirse enormemente este tiempo.

En conjunto, los experimentos realizados validan la eficacia de un enfoque híbrido para el análisis de vídeo deportivo, sentando una base sólida para futuras extensiones del sistema, como la mejora del seguimiento temporal, la optimización para tiempo real o su adaptación a otros deportes con estructuras geométricas similares.

References

- [1] Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsi-Ui Ik, and Wen-Chih Peng. Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. *CoRR*, abs/1907.03698, 2019. [2](#), [6](#)
- [2] ML6. Improving tennis court line detection with machine learning. <https://www.ml6.eu/en/blog/improving-tennis-court-line-detection-with-machine-learning>, 2021. Online blog article. [2](#), [3](#)
- [3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. [2](#)
- [4] Abdullah Tarek. Tennis analysis: Player, ball and court detection. https://github.com/abdullahtarek/tennis_analysis, 2022. GitHub repository. [2](#)
- [5] Sergey Yastrebkov. Tennis court detector. <https://github.com/yastrebksv/TennisCourtDetector>, 2021. GitHub repository. [2](#)