CIDAI Centre of Innovation for Data tech and Artificial Intelligence



Exploració de la detecció de dades biomètriques en partits de futbol



CIDAI-PAI-2024-06

Drets reservats. Aquest treball està disponible sota la llicència *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)*.

Segons els termes d'aquesta llicència, es pot copiar, redistribuir i adaptar l'obra amb fins no comercials, sempre que l'obra sigui citada adequadament, tal com s'indica a continuació.

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI.
Si s'adapta l'obra, s'ha de llicenciar amb la mateixa llicència *Creative Commons* o equivalent.

Si es crea una traducció d'aquest treball, s'ha d'afegir la següent exempció de responsabilitat juntament amb la cita suggerida: "Aquesta traducció no la va crear el *Centre of Innovation for Data tech and Artificial Intelligence* (CIDAI). CIDAI no es fa responsable del contingut ni de l'exactitud d'aquesta traducció. L'edició original en català serà l'edició autèntica i vinculant".

Qualsevol mediació relacionada amb disputes derivades de la llicència es durà a terme d'acord amb les normes de mediació de la World Intellectual Property Organization.

Citació suggerida: CIDAI-PAI-2024-06 // Exploració de la detecció de dades biomètriques en partís de futbol. CIDAI, 2024. Llicència: CC BY-NC-SA 4.0

Autors:

Coen Antens, Edgar Gracia Larrosa, Alex Rodríguez Marín
Centre de Visió per Computador (CVC)

Rafael Redondo Tejedor
Eurecat

Luis Ferraz
Kognia Sports

Índex

1. PROJECTE D'ALT IMPACTE (PAI) AMB HUB23 4
2. ESTIMACIÓ DE LA POSTURA EN EL FUTBOL: PER QUÈ ÉS RELLEVANT? 6
3. ESTIMACIÓ DE LA POSTURA: DEFINICIÓ I L'ESTAT DE L'ART 9
4. PRIMERS EXPERIMENTS 12
5. OBJECT KEYPOINT SIMILARITY IN KEYPOINT DETECTION 15
6. EXPERIMENTS AMB SOCCERNET 17
7. EXPERIMENTS AMB WORLDPOSE 20
8. CONCLUSIONS 24

ANNEX: POSSIBLES VIES DE CONTINUACIÓ 26

Projecte d'Alt Impacte (PAI) amb Hub23

1



Hub23 és un clúster empresarial que té per objectiu connectar i unir empreses que poden aportar tecnologia i innovació aplicables a totes les facetes de l'esport. Un exemple tecnològic evident és l'ús de càmeres de vídeo en activitats i esdeveniments esportius. I si les imatges capturades per aquestes càmeres es poden analitzar de forma automàtica amb tècniques de visió per computador i altres eines basades en Intel·ligència Artificial l'aportació de valor es multiplica.

Hub23 ha valorat diverses càmeres automàtiques (és a dir, dispositius que incorporen el processament d'imatges) comercials, com per exemple Pixellot, VEO, Spiideo, BePro, Hudl o Trace, per al seu ús en l'àmbit esportiu, en particular en el cas del futbol. Fruit d'aquesta valoració Hub23 ha seleccionat càmeres VEO (https://www.veo.co/). Aquestes càmeres, en principi, fan la detecció de la posició dels jugadors en el terreny de joc, però cap d'elles encara ha explorat la detecció automàtica de dades biomètriques dels jugadors.

En aquest context, es va proposar el desenvolupament d'un projecte d'innovació empresarial en la modalitat de Projecte d'Alt Impacte del CIDAI, amb l'objectiu d'estudiar la detecció automàtica de les extremitats dels jugadors de futbol. El projecte ha estat desenvolupat pel Centre de Visió per Computador (CVC), membre del CIDAI amb el suport d'Eurecat i de l'empresa Kognia Sports.

Estimació de la postura en el futbol: per què és rellevant?



2

La postura és un element altament rellevant en l'esport, ja que constitueix un component fonamental del rendiment atlètic que influeix en aspectes com la generació de força, la prevenció de lesions o les interaccions socials. Un alineament biomecànic òptim és essencial perquè garanteix que la força es transmeti de manera eficient al llarg de la cadena cinètica, permetent una major potència amb un menor consum d'energia. Aquesta estabilitat postural també està estretament relacionada amb el risc de lesió, fet que fa que la seva avaluació sigui indispensable per detectar i corregir possibles anomalies.

Aquesta rellevància s'estén especialment als esports d'equip, on una anàlisi precisa de la posició corporal és una eina revolucionària per estudiar les interaccions socials i el comportament col·lectiu, ajudant a optimitzar el rendiment global i a analitzar les activitats coordinades i dinàmiques de grans grups.

La tecnologia d'Estimació de la Posició (Pose Estimation) proporciona la base tecnològica per convertir imatges de retransmissions en dades objectives i quantificables, permetent una anàlisi avançada en el futbol d'elit, ja que la postura influeix en el rendiment, la prevenció de lesions, la tècnica i l'estratègia.

2.1. Aplicacions principals de l'estimació de la postura en el futbol

Anàlisi tècnica i biomecànica

Aquesta tecnologia ofereix mètriques detallades essencials per al perfeccionament del moviment i la salut de l'atleta. L'anàlisi biomecànica permet als entrenadors estudiar amb precisió la mecànica de moviments específics —com una xutada o una passada— mesurant els angles articulars i les velocitats dels membres per optimitzar la tècnica i maximitzar l'eficiència.

Un alineament biomecànic òptim és clau per a una producció de força eficient i una millora del rendiment. L'estimació de la postura és essencial en l'anàlisi dels moviments de xut en el futbol, permetent identificar diferències entre tipus de xut. En el cas dels porters, la detecció de la postura amb IA pot visualitzar la tècnica de volada, la posició i la resposta davant els xuts per assolir un rendiment d'elit. A més, aquesta anàlisi contribueix a la prevenció de lesions i al control de càrregues, ja que una detecció precisa pot identificar moviments de risc.

Intel·ligència tàctica i estratègica

L'estimació de la postura genera dades quantificables que informen la presa de decisions tàctiques i la planificació estratègica. La intel·ligència tàctica es deriva de les dades d'orientació corporal, on la direcció del cap, les espatlles i els malucs ofereixen informació clau sobre les intencions i possibles accions del jugador. Aquestes dades són essencials per modelar la viabilitat d'una passada i entendre les decisions dins del camp.

L'observació dels moviments del cap ajuda els analistes a determinar si el jugador ha avaluat la situació. A nivell d'equip, això permet una anàlisi més profunda de les estructures defensives i de les formacions. L'estimació de la postura també és una eina valuosa per analitzar les estratègies a pilota aturada, com els penals, estudiant com l'orientació corporal tant del xutador com del porter influeix en els seus processos de decisió.

Suport a l'arbitratge

L'estimació de la postura té un paper fonamental en la gestió dels partits, garantint una aplicació precisa del reglament. A més, en el suport a l'arbitratge, la detecció de la postura va més enllà de les decisions de fora de joc, ajudant en la identificació de faltes i altres situacions reglamentàries.

Detecció d'esdeveniments

La precisió i la riquesa en la detecció d'esdeveniments (com xuts, passades, serveis de banda, entrades o rematades) poden millorar-se notablement amb la integració de dades de posició corporal. Analitzant la configuració exacta del cos dels jugadors, es pot identificar no només que un esdeveniment ha tingut lloc, sinó també detalls crucials sobre la seva execució. Per exemple, saber quina part del cos ha tocat la pilota (peu, cap, pit) aporta un context valuós per comprendre millor la naturalesa d'una passada, un xut o un rebuig.

Models de valor

Els models de valor són instruments analítics essencials per quantificar l'impacte de diversos factors sobre un resultat determinat. Aquests models estableixen relacions quantitatives basades en dades, oferint informació sobre com diferents variables interactuen per influir en un resultat desitjat.

La integració de dades de posició és especialment útil per als models de valor avançats. Per exemple, els models moderns de Gols Esperats (xG) ja incorporen informació d'estimació de posició tant del xutador com dels defensors. Això millora de manera significativa la precisió en la predicció dels resultats dels partits, superant els models antics basats únicament en dades de posició.



Estimació de la postura: definició i l'estat de l'art

3



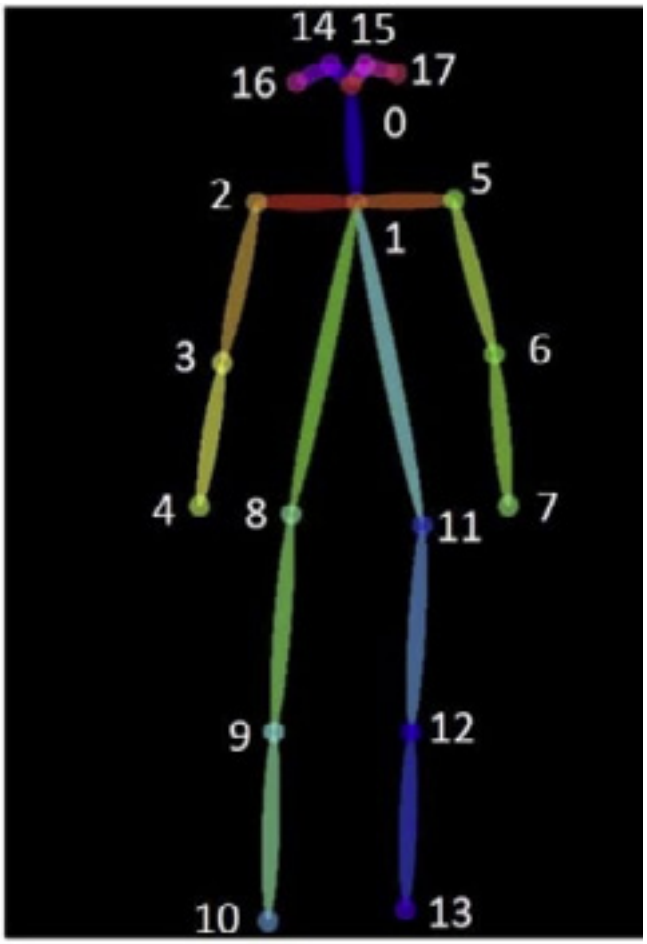
L'estimació de la postura (també coneguda per la denominació en anglès pose estimation) és una tasca fonamental dins la visió per computador i la intel·ligència artificial, que consisteix a detectar i seguir la posició i orientació de les diferents parts del cos humà en imatges o vídeos. Aquesta funcionalitat permet que les màquines interpretin i comprenguin la postura, el moviment i el comportament de les persones. L'estimació de la postura detecta i segueix punts clau semàntics tant en humans com en objectes. En el cas dels humans, aquests punts clau solen ser articulacions com l'espatlla dreta, el genoll esquerre, els colzes o els turmells. En objectes, es poden seguir punts específics com, per exemple, el llum de fre esquerre d'un vehicle.

El seguiment d'aquests punts clau en vídeos en temps real requereix molts recursos computacionals. Tradicionalment, això ha limitat l'aplicació pràctica d'aquests sistemes. Tot i això, els avenços recents en maquinari i l'eficiència dels models han fet possible una estimació de la postura en temps real de forma més assequible i efectiva. Actualment, els mètodes més eficients es basen en xarxes neuronals convolucionals (CNN), models dissenyats específicament per detectar i seguir punts clau amb gran precisió i rapidesa, i que són essencials en els sistemes d'estimació de la postura moderns.

Un dels estàndards més utilitzats per entrenar i avaluar models d'estimació de la postura és el conjunt de dades COCO (Common Objects in Context). Aquest defineix 17 punts clau al cos humà, cadascun anotat amb tres valors, (x, y, v), on x i y representen les coordenades del punt, i v indica la visibilitat (0: no anotat, 1: anotat però no visible, 2: visible). Els 17 punts clau del COCO inclouen: "nas", "ull esquerre", "ull dret", "orella esquerra", "orella dreta", "espatlla esquerra", "espatlla dreta", "colze esquerre", "colze dret", "canell esquerre", "canell dret", "maluc esquerre", "maluc dret", "genoll esquerre", "genoll dret", "turmell esquerre" i "turmell dret".

En la detecció tradicional d'objectes, els humans es representen normalment com a caixes delimitado-

Figura 1. Definició dels keypoints de COCO (Common Objects in Context)





res. Tot i que això pot ser útil, no permet entendre accions complexes. L'estimació de la postura va més enllà, ja que permet identificar la configuració específica de les parts del cos, i proporciona una comprensió més detallada del llenguatge corporal i del moviment. Aquesta capacitat és fonamental en múltiples aplicacions. En la conducció autònoma, per exemple, el seguiment en temps real de la postura permet predir millor el comportament dels vianants, millorant la seguretat i fent que els vehicles actuïn de manera més natural i humana. En l'àmbit esportiu, permet analitzar el rendiment, prevenir lesions i perfeccionar tècniques. En la salut, pot donar suport a la fisioteràpia, la monitorització de gent gran i el diagnòstic precoç de trastorns motors. Avui dia, molts accidents amb vehicles autònoms es produeixen per una manca de predicció del comportament humà: el vehicle autònom pot fer una aturada legal però inesperada, i un conductor humà pot xocar per darrere. Amb una millor comprensió de la postura humana, es poden anticipar millor les accions i adaptar la resposta de la màquina.

L'estimació de la postura humana es refereix específicament a la predicció de la configuració de les articulacions del cos humà en imatges o vídeos. Com que el moviment humà sol tenir un propòsit o estar relacionat amb accions específiques, conèixer la postura corporal és clau per reconèixer accions, interpretar gestos i entendre el contingut d'un vídeo. L'estimació de la postura en 2D se centra en determinar la ubicació espacial dels punts clau del cos en dues dimensions a partir d'imatges o vídeos. Els primers mètodes empraven tècniques manuals per descriure el cos com una figura d'esquelet. Aquestes aproximacions, però, tenien dificultats amb les oclusions, posicions complexes i escenes amb múltiples persones.

Els mètodes actuals basats en l'aprenentatge profund han millorat molt tant en rendiment com en robustesa. Les xarxes convolucionals s'han imposat com a enfocament dominant, entrenades amb grans conjunts de dades per detectar punts clau amb alta precisió. Alguns dels sistemes d'estimació de postura 2D més coneguts són OpenPose, CPN, AlphaPose i HRNet. A més, models més recents basats en transformers, com el SAPIENS de Meta, han ampliat els límits de l'estimació de la postura. SAPIENS aprofita l'aprenentatge a gran escala i arquitectures basades en Transformers per assolir una comprensió precisa i robusta de la postura humana, fins i tot en condicions complexes com ara oclusions, posicions inusuals o entorns amb molta informació visual. La seva capacitat de generalitzar bé en diferents escenaris el fa especialment valuós per a aplicacions del món real on la fiabilitat i l'adaptabilitat són fonamentals.

El mètode Slicing Aided Hyper Inference (SAHI) és una tècnica utilitzada en visió per computador per millorar el rendiment de la detecció d'objectes, especialment quan es treballa amb objectes petits o molt agrupats. Consisteix a dividir ("tallar") la imatge d'entrada en fragments més petits i aplicar la detecció sobre cada fragment per separat, en lloc de sobre la imatge sencera. SAHI és una estratègia d'inferència dissenyada per optimitzar el reconeixement quan els objectes són massa petits per ser detectats en una sola passada, quan hi ha molta densitat d'objectes, o quan els models perden qualitat per la necessitat de redimensionar imatges molt grans. És un mètode agnòstic respecte al model, cosa que vol dir que es pot aplicar a gairebé qualsevol detector d'objectes sense necessitat de tornar a entrenar-lo.

Primers experiments



4

Per iniciar l'estudi, Hub23 va proporcionar dos vídeos de partits de futbol i un vídeo addicional per il·lustrar els resultats que es volien obtenir (aquest darrer vídeo mostrava els resultats obtinguts per l'empresa Kognia Sports Intelligence). El CVC va dur a terme uns primers experiments per estudiar la viabilitat d'una solució basada en Intel·ligència Artificial, utilitzant models de Deep Learning (concretament, Detectron2).

Figura 2. Detecció de postura de jugadors amb Detectron2



La Figura 2 mostra els resultats obtinguts amb el model Detectron2. A partir d'aquestes proves inicials, es va concloure que tenia sentit continuar amb l'estudi. Seguidament, Hub23 va gestionar l'accés a unes càmeres VEO instal·lades al Club Esportiu Europa i va facilitar nous vídeos al CVC. El centre va dur a terme una revisió de l'estat de l'art en detecció de posicions humanes (Human Pose Detection) i va identificar el model Sapiens, desenvolupat per META, així com una tècnica anomenada Slicing Aided Hyper Inference (SAHI), pensada per aprofitar millor la resolució de les càmeres. Amb aquests nous algorismes es van realitzar els experiments que es comenten tot seguit.

Figura 3. Detecció de pose amb Detectro2: no detecta correctament la cama de la jugadora al quadre vermell.



A la Figura 3 es pot observar, en el quadre vermell gran (que correspon a una ampliació del quadre vermell petit), que el model Detectron2 no és capaç de determinar correctament la posició de les jugadores quan la resolució és massa baixa. En aquest context, el model Sapiens proporciona millors resultats, tal com es mostra a la Figura 4. Tanmateix, Sapiens no detecta ni la persona situada al costat del públic ni el porter. Finalment, la combinació del model Sapiens amb la tècnica de Slice Inference permet detectar la majoria d'aquests casos, com es pot veure a la Figura 5.

Figura 4. Detecció de pose amb Sapiens: no detecta la persona a la dreta del quadre vermell.



Els experiments realitzats fins ara han demostrat que la resolució de les imatges és un factor crític. El repte que planteja Hub23 és poder analitzar partits de futbol emprant una sola càmera, fet que implica que la resolució dels jugadors més allunyats pot arribar a ser molt limitada. Per estudiar els límits de cada mètode, es va dur a terme un experiment amb un únic jugador. La Figura 6 mostra els resultats obtinguts amb el model Detectron2, mentre que la Figura 7 presenta els resultats amb el model Sapiens de META. En aquests experiments, Detectron2 detecta correctament la posició del cos fins a una resolució de 66 píxels, mentre que les deteccions de Sapiens arriben fins als 43 píxels.

Figura 5. Detecció que combina Sapiens amb Slice Inference: deteccions són correctes.



Fins aquest punt, l'estudi s'ha centrat principalment en una anàlisi qualitativa dels resultats. Per tal d'avançar cap a una valoració més objectiva, a continuació s'inclouen dues seccions amb anàlisis quantitatives que permetran mesurar amb més precisió les prestacions de cada mètode. Abans, però, de comentar els experiments realitzats amb els dos nous conjunts de dades (datasets), cal definir un mètode que permeti comparar els resultats de manera rigorosa. Per aquest motiu, es presentarà primer el concepte de Object Keypoint Similarity (OKS), que servirà com a mètrica de referència per a l'avaluació dels models.

Figura 6. Experiments de resolució amb el model Detectron2



Object Keypoint Similarity in Keypoint Detection



En tasques de detecció de punts clau (keypoints), cal establir un mètode de comparació basat en la distància per tal d'avaluar la similitud entre els punts predits i els punts reals (Ground Truth). A diferència de les tècniques d'avaluació amb caixes delimitadores (que habitualment fan servir la mètrica Intersection over Union, IoU), els punts clau requereixen un enfocament més detallat i precís. Per aquest motiu, el conjunt de dades MS COCO va introduir la mètrica Object Keypoint Similarity (OKS), que permet quantificar la similitud entre dues configuracions de punts clau, tenint en compte l'escala de l'objecte i la variabilitat pròpia de cada punt, i que es considera l'estàndard en tasques de Pose Estimation. Aquesta mètrica és anàloga a la IoU però aplicada a punts clau, i es defineix com:

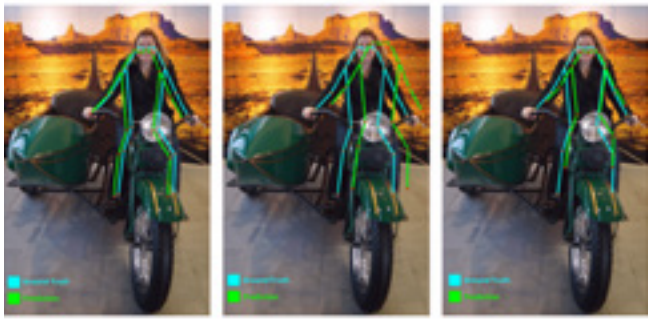
$$OKS = \frac{\sum_i^N \left[exp\left(-\frac{d_i^2}{2(sk_i)^2}\right) \cdot \delta(v_i > 0) \right]}{\sum_i^N \delta(v_i > 0)}$$

On per cada keypoint i:

- d: distància euclidiana entre el punt clau predit i el punt clau real per al punt i.
- s: escala de l'objecte real, definida com l'arrel quadrada de l'àrea segmentada de l'objecte de referència (GT).
- k: constant específica per al punt clau i, que ajusta la tolerància segons la seva importància o variabilitat.
- v: visibilitat del punt clau i; si v > 0, el punt és visible o ha estat anotat.
- (v>0): funció indicadora que val 1 si v > 0, i 0 en cas contrari.

Per ajudar a comprendre la fórmula de l'Object Keypoint Similarity (OKS), es poden observar els següents casos amb tres conjunts diferents de punts clau predits comparats amb els punts reals procedents dels Ground Truth (Figura 8). L'exemple s'ha extret d'aquesta pàgina web: https://learnopencv.com/object-keypoint-similarity/

Figura 8. Intuïció darrere de la fórmula OKS



- Primer model (esquerra): El conjunt de punts clau predits és molt proper als punts reals, indicant una alta similitud i, per tant, un valor OKS elevat.
- Segon model (mig): El conjunt de punts predits està força allunyat dels punts reals, la qual cosa implica una baixa similitud i un valor OKS baix.
- Tercer model (dreta): El conjunt de punts predits és proper als punts reals però amb una certa escala diferent (amplificat o reduït), afectant lleument la similitud i resultant en un valor OKS intermedi.

Aquests exemples il·lustren com OKS mesura la proximitat i l'adequació dels punts predits, tenint en compte no només la distància sinó també la mida relativa de l'objecte.

Experiments amb SoccerNet

6



Amb l'objectiu de continuar l'avaluació del model Sapiens de META en contextos més diversos, s'ha optat per utilitzar el dataset SoccerNet, que proporciona una àmplia varietat de partits amb diferents enfocaments, condicions lumíniques, climàtiques i moments del dia. Aquesta diversitat permet posar a prova la robustesa del model en escenaris més realistes i variats.

SoccerNet (www.soccer-net.org) és un conjunt de dades a gran escala dissenyat específicament per a la comprensió de vídeos de futbol. Al llarg dels anys, ha anat evolucionant per incloure diverses tasques relacionades amb aquest àmbit, com ara la detecció d'accions (Action Spotting), el calibratge de càmeres, la reidentificació i el seguiment de jugadors. El dataset està format per 550 partits complets retransmesos, així com per 12 partits enregistrats amb una única càmera, tots extrets de les principals lligues europees. Aquesta gran quantitat i varietat de dades permeten desenvolupar i avaluar models més robustos i generalitzables en l'anàlisi de futbol. Per aquests motius, SoccerNet representa un entorn idoni per desenvolupar i validar les nostres solucions d'anàlisi i detecció basades en vídeo en l'àmbit del futbol.

Per tal de simplificar l'experiment, es van seleccionar 10 partits del dataset, dels quals es van extreure al voltant de 140 imatges representatives. A partir d'aquestes, es va construir un Ground Truth manual per a 50 fotogrames, utilitzant l'eina CVAT (Computer Vision Annotation Tool). Les anotacions es van fer mitjançant un esquelet basat en els 17 punts clau del format COCO.

Figura 9. Comparacions utilitzant bounding box.



Inicialment, l'àrea s es calculava mitjançant una Bounding Box que envoltava tots els punts clau tal i com es veu a la Figura 9. No obstant això, aquest mètode sovint generava àrees molt grans o molt petites, especialment en casos on els punts clau estaven dispersos o mal alineats, cosa que afectava negativament la mètrica OKS.

Figura 10. Comparacions utilitzant Convex Hull.



Per millorar aquesta estimació, es va optar per calcular l'àrea mitjançant el convex hull dels punts clau com es veu a la Figura 10. El convex hull és la figura geomètrica convexa més petita que conté tots els punts clau, i s'assembla a una goma elàstica estirada al seu voltant. Aquest enfocament proporciona una estimació més realista de l'àrea real ocupada pel cos de la persona, perquè:

- Redueix la influència de punts clau dispersos o erronis
- Evita àrees desproporcionades causades per Bounding Boxes rectangulars
- Genera valors d'àrea més estables i propers a la mitjana real.

Figura 11. Millor càlcul OKS.



Per complementar l'anàlisi quantitativa, s'han seleccionat dos exemples representatius:

- Millor cas: una imatge on el model Sapiens detecta correctament tots els punts clau, amb una puntuació OKS molt alta, propera a 1.00. Aquesta situació es dona en condicions ideals: bona il·luminació, alta resolució i una perspectiva favorable de la càmera. A la Figura 11 es pot veure el jugador situat a la dreta, que ha estat ampliat per mostrar amb detall la qualitat de la detecció.

- Pitjor cas: una imatge amb una puntuació OKS de 0.07, deguda a una combinació de baixa resolució, moviment ràpid i perspectiva de la càmera. Quan la càmera està més a prop del camp, els jugadors ocupen més píxels a la imatge, fet que millora la qualitat de les prediccions. En canvi, en aquest cas concret, la distància i l'angle de visió han dificultat la detecció precisa. Diversos punts clau apareixen significativament desplaçats, tal com es pot observar a la Figura 12, on s'ha ampliat el jugador que ha estat detectat incorrectament.

Figura 12. Pitjor càlcul OKS.



Experiments amb WorldPose

7



Curat de dades

WorldPose és un dataset recent i innovador destinat a l'estimació global de posicions corporals de múltiples persones en entorns reals i complexos, basat en imatges capturades durant la Copa Mundial de la FIFA 2022. Aquest dataset destaca per la seva configuració multivista, amb múltiples càmeres fixes i mòbils distribuïdes en diferents estadis, cosa que permet obtenir dades en 3D amb gran precisió sobre el moviment dels jugadors en una àrea extensa. WorldPose ofereix aproximadament 2,5 milions de posicions en 3D distribuïdes en més de 80 seqüències, així com informació detallada sobre el calibratge de càmeres i els vídeos associats. Aquesta riquesa de dades fa que sigui especialment valuós per a l'anàlisi de la posició i el moviment dels jugadors en partits de futbol. Als experiments, s'ha utilitzat WorldPose com a referència per validar i comparar els models d'estimació de posició corporal, aprofitant la seva qualitat i complexitat per posar a prova la robustesa i precisió dels nostres mètodes.

Aquest dataset proporciona, per a cada fotograma, fins a 45 keypoints per a cada jugador detectat. Tot i això, per tal de mantenir la coherència amb les anàlisis anteriors i adequar-nos a les necessitats concretes del projecte (basat en el format COCO), només s'han utilitzat els 17 keypoints habituals.

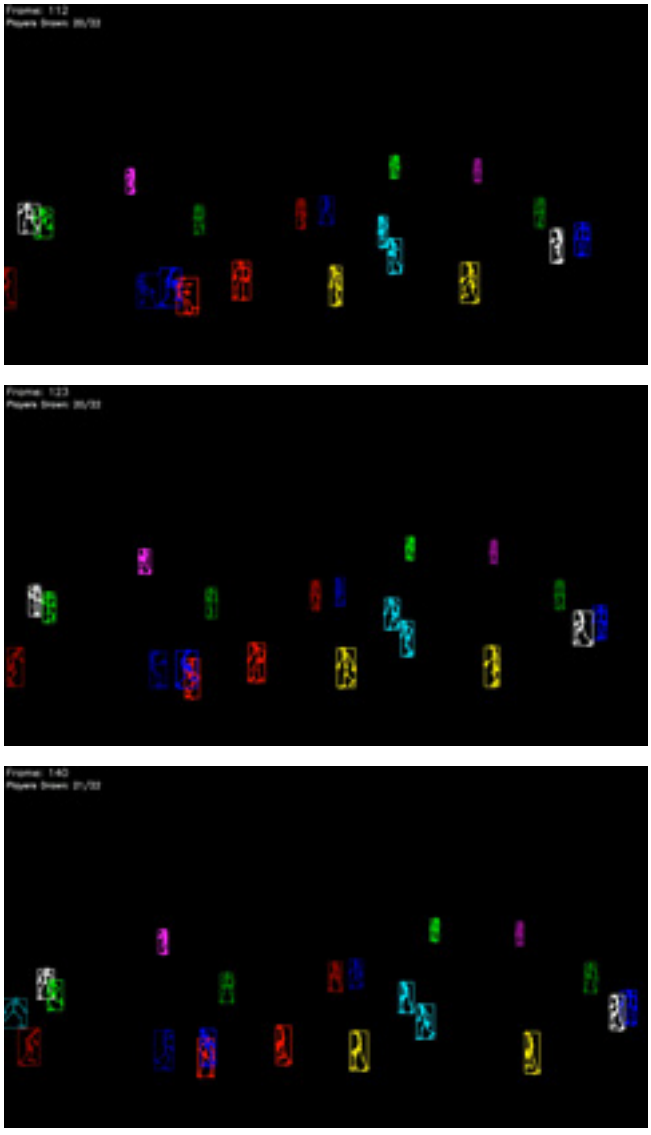
Aquesta estratègia ha permès utilitzar les anotacions de WorldPose com a Ground Truth (GT) automàtic, estalviant així la tasca d'etiquetar manualment cada imatge, alhora que proporciona una base sòlida per continuar validant el rendiment del model.

Per a aquesta tasca Eurecat va desenvolupar scripts per formar un flux de treball complet per a la visualització i anotació de moviments dels humans 3D (SMPL[1]) projectats a imatge 2D, utilitzable per entrenar o verificar els models de visió per computador per la detecció de jugadors en el context FIFA. El codi desenvolupat genera vídeos amb esquelets SMPL projectats sobre imatges buides (o amb fons opcional), a partir de les posicions i paràmetres de càmera. En particular projecta les articulacions 3D del model SMPL a coordenades 2D d'imatge, creant anotacions amb les posicions 2D conjuntament amb les caixes delimitadores (bounding boxes).

Les dades es van processar en format JSON amb la descripció dels punts de control formats pels esquelets i caixes delimitadores. Entre d'altres el canvi de format va necessitar extreure les parts essencials del models en format Pickle[2] per a ús amb PyTorch o smplx. En particular, la forma base del cos, direccions de deformació segons paràmetres de forma, direccions de deformació segons la posició, regressor d'articulacions, pesos d'influència, jerarquia d'ossos i cares de la mal.

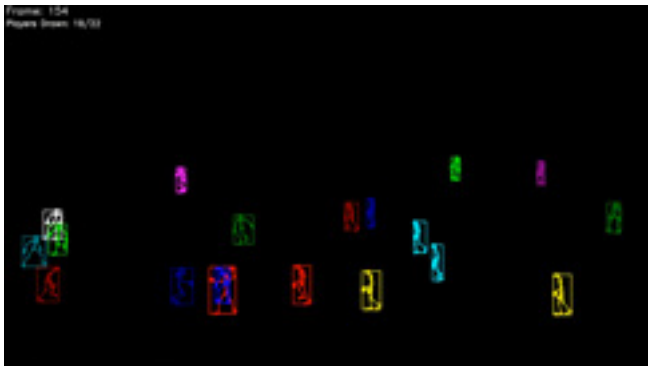
Per la verificació del procés de conversió del format 3D a 2D es va processar del ordre de 90 vídeos d'entre 30-60 segons. La Figura 13 il·lustra alguns del quadres processats.

Figura 13. Exemple de 4 quadres (112, 123, 140, 154, no consecutius) de la mateixa seqüencia de vídeo processada del conjunt de dades WorldPose FIFA on es mostra la projecció 2D dels models 3D SMPL dels jugadors.



1 https://smpl.is.tue.mpg.de

2 https://docs.python.org/3/library/pickle.html



Validació amb WorldPose

Per aquesta avaluació s'han seleccionat 2 vídeos del dataset WorldPose, dels quals s'han extret 25 fotogrames per vídeo, sumant un total de 50 fotogrames analitzats. Aquest conjunt de dades ha estat utilitzat per comparar les prediccions del model Sapiens amb les anotacions automàtiques existents.

En total, es van obtenir:

- 418 + 419 = 837 parelles de anotacions (GT) i prediccions. Amb aquestes s'han fet els càlculs de la mètrica OKS com a la secció anterior.
- 106 + 107 = 213 anotacions (GT) sense predicció per part del model, respectivament, en cadascun dels vídeos.
- 25 + 10 = 35 prediccions sense GT en cadascun dels vídeos. Aquestes prediccions han estat majoritàriament descartades de l'anàlisi quantitativa ja que corresponien a persones no rellevants pel nostre estudi (entrenadors, àrbitres, observadors...).

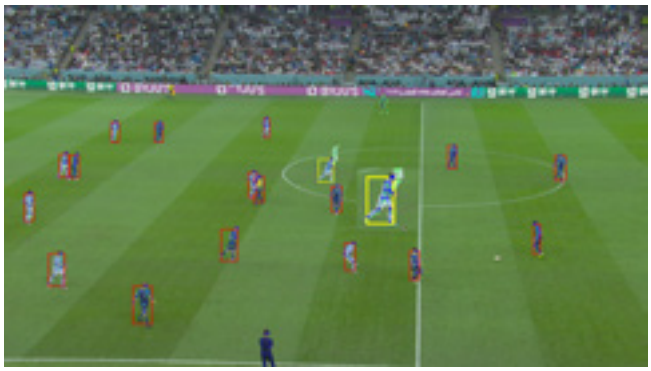
Tot i que les anotacions del WorldPose poden contenir errors puntuals —atès que són generades per models d'IA i no manualment—, la seva qualitat general és suficient per permetre una avaluació objectiva i pràctica del model. Com que el nostre interès se centra únicament en jugadors, la comparació entre GT i prediccions s'ha realitzat filtrant els altres subjectes detectats.

Al igual que amb SoccerNet, s'ha utilitzat la mètrica Object Keypoint Similarity (OKS) per quantificar la precisió de les deteccions. Tot i la naturalesa automàtica del GT, aquesta mètrica segueix sent útil per avaluar la consistència del model respecte a un estàndard conegut i aplicable a dades reals.

Com a complement a l'avaluació quantitativa, s'han identificat 20 casos, els 10 millors i pitjors OKS, mostrem dos casos representatius:

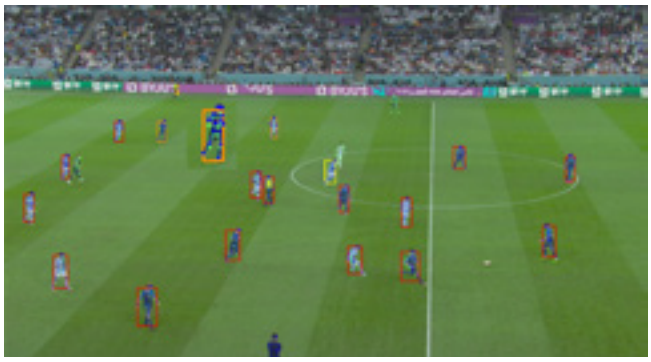
Millor cas – Figura 14: Detenció precisa de tots els punts clau amb una puntuació OKS elevada. Les condicions visuals del fotograma (resolució i angle) han estat òptimes i el model ha funcionat amb una gran precisió.

Figura 14. Millor càlcul OKS.



Pitjor cas - Figura 15: Aquest és un mal cas amb OKS baix perquè els punts predits no coincideixen amb els del ground truth i, a més, l'àrea bbox (bounding box) del jugador és molt petita, fet que amplifica l'error en la mètrica.

Figura 15. Pitjor càlcul OKS.



A continuació es mostren tres casos representatius on el model Sapiens no detecta correctament els jugadors. En tots els casos es visualitzen les anotacions del ground truth (en verd) proporcionades per WorldPose, les prediccions del model (en blau) i la bounding box associada a cada detecció (en vermell).

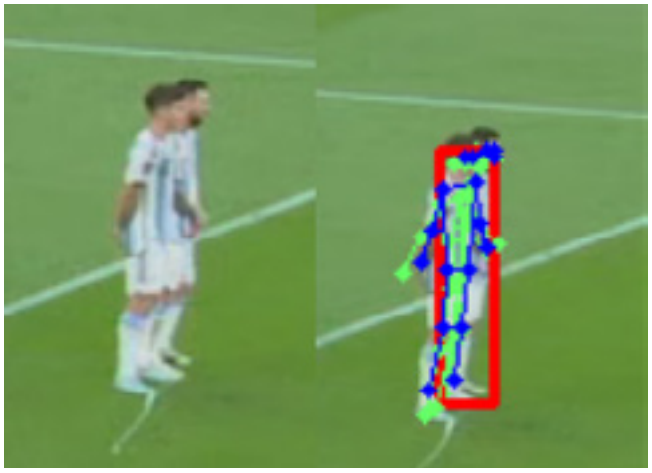
Figura 16. Oclusió temporal.



Cas 1: Oclusió temporal (jugador i àrbitre creuant-se)

En aquest cas (Figura 16) es mostren tres fotogrames consecutius on es pot observar com el model deixa de detectar el jugador just quan es creua amb l'àrbitre. El GT, però, segueix present i ben posicionat en verd. Aquesta situació posa de manifest que el model no gestiona bé les oclusions breus i perd deteccions en escenaris dinàmics.

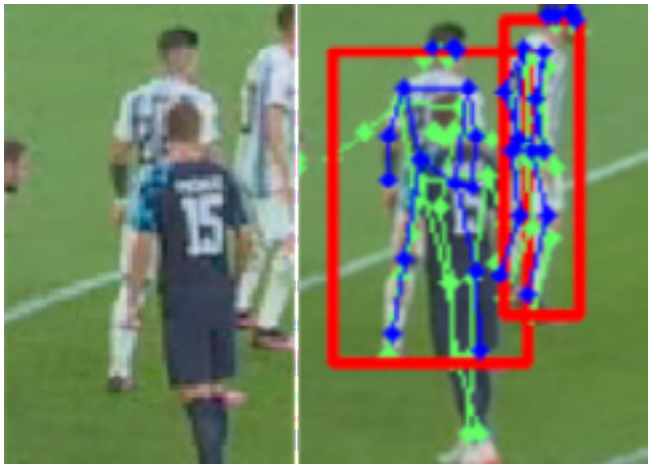
Figura 17. Confusió de deteccions.



Cas 2: Dos jugadors molt junts (confusió de deteccions)

La Figura 17 mostra un cas on dos jugadors estan molt propers, i el GT proporciona anotacions per a tots dos. Tanmateix, el model només genera una sola predicció, probablement perquè els cossos es solapen i la detecció es confon. Això redueix el nombre de deteccions.

Figura 18. Error de fusió de deteccions.



Cas 3: Error de fusió de deteccions

En aquest cas (Figura 18), el model detecta només una persona però assigna punts clau de dos jugadors diferents a la mateixa predicció, generant una detecció incorrecta que barreja informació dels dos cossos.

Dataset	Nº frames	Jugadors amb GT	Prediccions correctes*	Jugadors sense detecció	Prediccions sense GT
SoccerNet	50	686	575	106	27
WorldPose	50	1050	448	213	35

* SoccerNet -> Prediccions que coincideixen amb el GT amb oks > 0.5 (match correcte).

WorldPose -> Prediccions que coincideixen amb el GT amb oks > 0.3 (match correcte).

Hem decidit baixar aquest valor llindar a WorldPose ja que els valors oks a aquest dataset son per definició més baixos degut a que al dataset SoccerNet el GT s'ha fet a partir de les prediccions de Sapiens i sovint les prediccions que han sigut bastant correctes no han estat corregides (OKS = 1), mentre que a WorldPose el GT ha sigut proporcionat.

Conclusions

8

A partir dels experiments que s'han fet amb els vídeos que Hub23 va facilitar al Centre de Visió per Computador es poden treure algunes conclusions preliminars:

- El model Detectron2 detecta correctament un jugador des de 1000 fins a 66 píxels mentre el Sapiens fa una detecció correcta des de 1000 fins a 43 píxels.
- Segons les mètriques oficials amb el dataset de COCO (17 Keypoints) el Average Precision del Detectron2 (X101-FPN) es 66.0 i el Average Precision de Sapiens (0.3B) es 79.6.
- El temps d'inferència del Detectron2 fa que el model gairebé pugui funcionar en temps real (múltiple frames/segon, depèn del backbone).
- El temps d'inferència del Sapiens es qüestió de segons i fa que el model, basat en Transformer, no pugui funcionar en temps real.

Amb l'objectiu de validar la robustesa del model en escenaris més diversos, es va ampliar l'estudi utilitzant el dataset SoccerNet. Es van anotar manualment 50 fotogrames mitjançant CVAT, i es va aplicar la mètrica Object Keypoint Similarity (OKS) per avaluar el rendiment. Inicialment, l'àrea de referència es calculava amb bounding boxes, però aquest mètode generava valors inconsistents. L'ús del convex hull dels punts clau va permetre obtenir àrees més estables i realistes, millorant la fiabilitat de la mètrica.

L'ús del dataset WorldPose ha estat fonamental per continuar l'avaluació del nostre model de manera eficient, sense necessitat d'anotar dades manualment. El balanç general demostra que podem continuar utilitzant aquest recurs per millorar i ajustar el model, estalviant recursos i centrant-nos en casos més difícils o crítics.

En conjunt, els experiments mostren que el model Sapiens, especialment quan es combina amb tècniques com SAHI i es valida amb mètriques com OKS, és una eina prometedora per a la detecció de dades biomètriques en partits de futbol. Tot i això, cal seguir treballant per millorar-ne l'eficiència i la capacitat de detecció en situacions on els jugadors apareixen molt petits o amb moviment ràpid, com passa sovint quan la càmera està lluny del joc i la resolució efectiva és baixa.

Annex: possibles vies de continuació



Robustesa del sistema davant de condicions adverses

Avaluar el rendiment del detector en situacions com boira, pluja, ombres i partits nocturns:

- Boira: La baixa visibilitat afecta la precisió del detector.
- Pluja: Gotes d'aigua o reflexos poden generar soroll.
- Ombres: Canvis bruscos d'il·luminació o zones fosques.
- Partits nocturns: Menor il·luminació, possibles problemes amb contrast i qualitat de la imatge.

Estudi de l'anàlisi de coherència temporal entre fotogrames

Millorar l'estabilitat del detector utilitzant la coherència temporal. Això implica aprofitar el flux d'informació entre fotogrames per evitar salts o inconsistències en la detecció de l'esquelet.

Possibles tasques: Implementar heurístiques com Optical Flow o filtres de Kalman.

Seguiment i estimació de postures dinàmiques al llarg del temps

Alguns del les postures requereixen un anàlisis del cos al llarg de temps en una seqüència de fotogrames. Això implicaria el desplegament de sistemes de seguiment (tracking) de persones a sobre de la detecció de postures.

Robustesa del sistema en situacions allunyades de la càmera

A mesura que el jugador s'allunya de la càmera la manca de resolució de la imatge es fa més evident i la capacitat de detecció de la postura del jugador és menys precisa.

- Estudi de limitacions relacionades amb la manca de resolució.
- En base a l'estudi, metodologies per enfortir la manca de resolució en imatges de persones/jugadors.

Estimació de poses en 3D

En base a les necessitats de mesures biomètriques (encara no descrites) pot ser rellevant l'ús de models de estimació de poses en 3D que habiliti la mesura de postures més acurades.

Gestió de oclusions

En aquest context les oclusions de parts del cos poden ser degudes a altres jugadors (arbitre) que interfereixen la vista entre càmera i jugador o bé el mateix jugador que per la seva postura no deixa veure totes les extremitats a l'hora.

- Desenvolupament de algorismes per gestionar extremitats oclosos pel propi jugador.
- Desenvolupament de algoritmes temporals per gestionar oclusions entre jugadors.

Nota: potser aquest es pot dividir en 2 reptes perquè les tècniques son diferents.

Calibratge de la perspectiva del camp

Segons quines mètriques relacionades amb la velocitat serà necessari el calibratge de les càmeres i estimació de la localització del jugador en el camp per tal de extreure mètriques de velocitat en fiables en qualsevol punt del camp.

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

C/Bilbao, 72 Edif. A.
08005 - Barcelona
Tel. +34 93 7419 100
info@cidai-catalonia-ai.eu
www.cidai.eu

Promotors

Membres