PN-III-P2-2.1-PED-2019-4995

PETRA

Detecția și urmărirea persoanelor pentru roboți sociali și mașini autonome

(People detection and **tra**cking for social robots and autonomous cars)

Etapa I - 2020

Raport științific și tehnic

Cuprins

1. Introducere

Detecția, urmărirea și recunoașterea persoanelor este o capacitate valoroasă pentru aplicațiile de vedere computerizată. Cu toate acestea, aceste sarcini sunt destul de dificil de realizat în mod autonom și, deși s-au obținut rezultate semnificative, acestea sunt încă o provocare tehnologică majoră. Urmărirea persoanelor, spre deosebire de alte sarcini de recunoaștere și interpretare, este dificilă atât din punctul de vedere al recunoașterii și predicției traiectoriei, cât și din cel al identificării instanțelor reale. Persoane parțial vizibile, reflexii în oglinzi sau ferestre și obiecte care seamănă foarte mult între ele, toate impun ambiguități intrinseci, astfel încât chiar și oamenii ar putea să nu fie de acord cu o aceeași soluție. Mai mult, stabilirea unor metrici de evaluare cu parametri liberi și definiții ambigue (cum se întâlnește des în literatură) duc adesea la rezultate cantitative contradictorii.

Obiectivul principal al proiectului PETRA este dezvoltarea unei platforme software pentru dezvoltarea de aplicații care necesită detecția și urmărirea persoanelor în medii reale. Scopul este acela de a proiecta și implementa o platformă și un set de algoritmi asociați care să poată fi ușor utilizați și integrați în aplicații care necesită atât detecția și urmărirea persoanelor de către roboții sociali în spații închise, cât și efectuarea aceleiași sarcini în spații deschise, în particular detecția și urmărirea pietonilor.

Obiectivul Etapei I (2020) a proiectului a fost acela de a realiza proiectarea algoritmilor care vor fi utilizați în platformă, investigarea seturilor de date utilizate în antrenare, colectarea de seturi de date noi.

2. Detecția persoanelor

Detecția persoanelor reprezintă o provocare curentă pentru domeniile inteligenței artificiale, ale învățării automate și pentru vizualizare asistată de calculator (computer vision). Avantajele unui sistem care să detecteze persoane sunt evidente – pornind de la sisteme de supraveghere pentru magazine, case, etc și continuând cu sisteme care să ajute persoanele cu nevoi speciale, dar și sisteme care ajută mașinile autonome să conducă singure, detecția obiectelor în general, dar și a pietonilor în mod particular, fiind una dintre cele mai importante probleme existente în cadrul mașinilor autonome.

Din punct de vedere al clasificării tipurilor de rețele de detecție de oameni, putem discuta despre rețele care detectează persoane în cadrul unor încăperi (utile, de exemplu, pentru roboți care ajută persoanele la diverse activități), și despre rețele care clasifică pietonii în afara unei încăperi, de exemplu pe stradă sau în trafic.

Pentru ambele tipuri de probleme sunt utile și arhitecturile care se ocupă de ”urmărirea” persoanelor, în sensul de a detecta că în imagini succesive obținute dintr-un videoclip este prezentă aceeași persoană sau nu. Este necesar de multe ori să verificăm dacă într-un videoclip este vorba de aceeași persoană sau dacă o altă persoană a intrat în scenă. Un exemplu ar fi un robot care ajută o persoană nevăzătoare, și trebuie să identifice constant persoana pe care trebuie să o ajute, dar și persoanele din jurul acesteia. În cadrul proiectului PETRA, toate aceste aspecte sunt luate în considerare.

2.1 Analiza celor mai bune arhitecturi pentru detecție de persoane

În ordine istorică, primele rețele care au apărut în cadrul detecției de obiecte în general (și persoane în particular) constau în două stagii. În 2016, R-CNN [1] a fost primul detector de obiecte în 2 faze. Într-o rețea cu două stagii/ etape, primul pas constă în extragerea unor regiuni din imagine, care vor fi folosite mai departe în stagiul al doilea ca să clasifice obiecte din acele regiuni. R-CNN folosește algoritmul de căutare selectivă [2] pentru a genera 2000 de regiuni care sunt propuse spre detecție.

Chiar dacă algoritmul merge bine din punct de vedere al rezultatelor, timpul de antrenare durează destul de mult, pentru că sunt 2000 de regiuni propuse pentru fiecare imagine. De asemenea, timpul de detecție pentru o imagine durează aproape un minut, nefiind o soluție viabilă pentru un algoritm care vrea să se execute în timp real. Algoritmul de căutare selectivă este un algoritm static, care va produce mereu aceleași regiuni, ceea ce ar putea duce la diverse greșeli.

Din aceste motive, autorii arhitecturii R-CNN au venit cu o versiune mai bună, Fast-RCNN [3]. Diferența față de versiunea precedentă constă în faptul că în R-CNN fiecare regiune trebuie să fie folosită ca intrare pentru o rețea de tip CNN diferită, deci erau 2000 de rețele care se executau în paralel. Acest lucru nu mai este valabil în cazul Fast R-CNN, deoarece rețeaua de tip CNN primește toată imaginea ca intrare și generează o hartă caracteristică convoluțională pentru toată imaginea.

Ultima îmbunătățire adusă acestei arhitecturi este facută în [4] și se numește Faster R-CNN (R-CNN și mai rapid). Cel mai mare avantaj al acestei arhitecturi este că introduce o rețea separată, special pentru a propune regiuni (Region Proposal Network), în loc de a se uitliza algoritmul de căutare selectivă. Faster R-CNN include conceptul de ancore pentru rețeaua RPN – în general obiectele tind să aibă aceleași rapoarte între dimensiuni, dacă ne gândim la oameni acest aspect este evident.

Pe lângă această rețea, există și altele care se folosesc la detecția de obiecte în 2 stagii, de exemplu R-CFN. [5], dar au fost depășite ca performannță, în special legat de viteză, de rețelele cu o singură etapă de procesare (single stage). Cea mai cunoscută rețea de acest tip este YOLO [6] , care a apărut în mai multe versiuni incrementale, ultima folosită fiind YOLO v3[7].

Următoarea arhiectură analizată este SSD [8]. Ca și YOLO, operează într-un singur pas, spre deosebire de R-CNN. Poate fi folosită în aplicații reale, având viteze suficient de bune, de până la 60 de imagini detectate pe secundă, pentru modelul standard, respectiv pană la 22 pentru modelul îmbunătățit cu precizie mai bună. Imaginea este pusă în straturi convoluționale, obținând o hartă de caracteristici. Pentru fiecare locație, sunt calculate anumite coordonate în care pot fi obiecte, pentru fiecare se calculează un scor de încredere și 4 coordonate pentru dreptunghiul care încadrează obiecte. Diferența față de YOLO este că se fac mult mai multe detecții, folosind un algoritm care lucrează mu dimensiuni multiple, fiind posibil un număr maxim de 8722 de obiecte detectate. SSD folosește ancore, de asemenea, folosesc convoluții dilatate, iar varianta îmbunătățită folosește imagini cu rezoluții mai mari.

O altă rețea care merge într-o singură etapă este RetinaNet [9], și este bazată pe arhitectura folosită de FPN [10]. Arhitectura de tip FPN este oarecum similară cu SSD – imagini cu diverse dimensiuni sunt folosite pentru a calcula o hartă de caracteristici. În loc să existe o singură ieșire la finalul convoluțiilor, pentru fiecare reducere dimensională se calculează o hartă de caracteristici, sunt făcute detecții pentru fiecare dimensiune, iar la final se reduc duplicatele. Pentru a avea rezultate mai bune, au construit o arhitectură specială ca să aibă caracteristici bune pentru dimensiuni mici și mari. RetinaNet folosește ancore, de asemenea, ca toate arhitecturile prezentate până acum. Față de alte arhitecturi, ei folosesc o funcție de penalizare diferită, care face ca penalizarea pentru obiectele foarte bine clasificate să fie minimală. De asemenea, au o subrețea care folosește Resnet pentru detecție de obiecte, o subrețea pentru clasificare, și de asemenea o subrețea pentru regresia locațiilor finale ale obiectelor.

Ultima categorie pe care o discutăm constă în rețele care nu mai folosesc deloc conceptul de ancore descris anterior. Un exemplu este rețeaua CornerNet [11], care folosește o rețea de tip clepsidră pentru extragerea de caracteristici. Ei elimină ancorele și reprezintă obiectele ca având 4 coordonate, din care se prezic doar 2 – colțul stânga sus și dreapta jos. Există un modul de predicție care include un strat pentru colțuri. Acest strat prezice o hartă cu probabilități pentru colțuri, și pentru fiecare colț

se calculează un vector caracteristic care încearcă să minimizeze distanța dintre colțuri pentru un obiect. Un obiect este prezis dacă distanța dintre cele două colțuri este mai mică decât o limită setată. De asemenea se folosesc tehnici de suprimare pentru a elimina locațiile multiple pentru un singur colț. Antrenarea se folosește folosind imagini în care obiectele sunt adnotate cu dreptunghiuri, din care se extrag doar colțurile. De asemenea, se încearcă să se obțină vectori similari pentru colțurile care aparțin aceluiași obiect. Există diverse implementări, cum ar fi CornerNet Lite [12], care este mai rapid, dar obține rezultate puțin mai slabe. Cu toate acestea, timpul nu este mai bun decât la YOLO.

Ultima rețea discutată este ExtremeNet [13], care pornește de la CornerNet, dar încearcă să găsească toate cele 4 colțuri și în plus și mijlocul obiectului. De asemenea, folosesc noțiunea de punct extrem, care poate fi diferită de un colț în sens tradițional, putând fi chiar în afara obiectului în sine.

La momentul de față, pentru partea de urmărire există diverse rețele neurale, care acționează diferit. De exemplu GOTURN [14] este o rețea care are rezultate foarte bune, până la 100 de cadre pe secundă, dar testarea se face offline, cunoscând și imaginile ulterioare imaginii analizate.

O altă arhitectură este ROLO [15], care se bazează pe YOLO, folosește o rețea de tip CNN și alta de tip LSTM și poate detecta și urmări obiectul chiar dacă există ocluzii, inconveniente vizuale, etc, dar problema este că funcționează doar pentru un singur obiect, astfel încât dacă nu este îmbunătățită pentru a detecta și urmări mai multe obiecte sau persoane, nu poate fi utilizată în aplicații mai generice.

Există și rețele care acționeaza online obiecte multiple, cum ar fi DeepSort [16], care este antrenată special pentru urmărire de persoane și poate fi utilizată cu rezultate bune pentru studiul de față. Folosește un filtru Kalman pentru a face predicții noi folosind obiecte existente, iar pentru a asocia obiectele dintr-o imagine cu una precedentă se folosește algoritmul ungar.

2.2 Propunere de arhitectură pentru detecție de persoane

Plecând de la modelele prezentate mai sus, propunerea noastră pentru proiectul curent este optimizarea a două arhitecturi existente astfel încât să obținem rezultate cât mai bune pentru scopul nostru – detecția pietonilor, arhitectură ce va ajuta inclusiv la urmărirea și identificarea acestora de la un cadru la altul.

Prima propunere constă în modificarea arhitecturii YOLO v3 – modelul cel mai folosit pentru detecția obiectelor, în general, astfel încât să se obțină o rețea mai specifică pentru ceea ce ne dorim noi.

Primul pas constă în modificarea ancorelor folosite la detecția obiectelor, astfel încât să rămână doar ancore necesare detecției de pietoni – în loc de 9 ancore cât sunt folosite în mod tradițional, vom păstra doar 2, una pentru adulți și alta pentru copii, aceștia având rapoartele puțin diferite față de adulți, fiind semnificativ mai mici ca înălțime.

A doua modificare a rețelei constă în procedeul de ”ajustare fină” (fine tuning), ce presupune eliminarea ultimelor straturi ale rețelei, în care se clasifică obiectele în mai multe categorii, păstrând doar 2 categorii – persoană sau alt obiect. Rețeaua va fi antrenată de la zero folosind seturile noastre de date înregistrate în campusul facultății de Automatică și Calculatoare, cât și seturi de date clasice.

Cea de-a doua propunere de rețea pleacă de la CenterNet, una dintre cele mai folosite rețele fără ancore. La CornerNet se încearcă detecția celor două colțuri extreme și centrul imaginii, iar noi vom modifica rețeaua astfel încât să mai adăugăm 2 straturi – unul pentru detecția lățimii și altul pentru detecția înălțimii. În acest fel, dacă s-au detectat cele 2 puncte extreme, pe lângă corecția pe care o face CenterNet legată de punctul din centru, vom face de asemenea predicția lățimii și înălțimii

persoanelor, astfel încât dacă diferențele sunt mari între predicția lățimii și lățimea calculată din cele două colțuri, se vor face ajustări care țin cont de ambele rezultate.

De asemenea, această rețea va fi antrenată de către noi folosind doar 2 tipuri de obiecte (persoană și alt obiect), folosind seturile de date din cadrul campusului, dar și seturi de date tradiționale.

3. Urmărirea persoanelor

3.1 Soluții pentru predicția traiectoriei

Predicția traiectoriei este una din abordările recente ce îmbunătățește urmărirea de persoane. Identitățile persoanelor detectate într-un cadru pot fi mai precis identificate în cadrele următoare prin prezicerea pozițiilor viitoare bazate pe observații anterioare. Lucrarea *Tracking by prediction* [17] propune un sistem complet de urmărire de persoane într-o secvență de imagini. Acest sistem integrează atât soluții de detecție, cât și de urmărire de obiecte. Detecția persoanelor este realizată folosind o rețea generativă adversarială, ce calculează probabilitatea fiecărui pixel de a aparține unui om. Modelul generativ incorporează un sistem encoder-decoder ce combină straturile convoluționale cu o rețea *Long Short-Term Memory* (LSTM). Distribuția de probabilități ce este returnată de către modelul discriminativ este trimisă la intrarea rețelei LSTM, care realizează clasificarea finală bazată pe hărțile de probabilitate din cadrele precedente. Pentru procesul de urmărire, sistemul folosește o plajă de detecții ce sunt asociate cu noile detecții prin predicția traiectoriei. Designul arhitecturii ajută în cazul ocluziilor mari în așa fel încât chiar dacă o persoană nu a mai putut fi detectată pentru un număr de cadre, aceasta poate fi re-identificată.

Lucrarea de pionierat realizată în *Social LSTM* [18] este punctul de plecare pentru multe direcții de cercetare în cadrul subiectului de prezicere a traiectoriei persoanelor. Provocarea este prezicerea simultană a mai multor traiectorii precise a persoanelor detectate într-o secvență de imagini RGB. Soluția propusă asociază fiecărui om detectat o rețea LSTM ce prezice pozițiile viitoare. Fiecare LSTM este conectat la un strat de pooling ce partajează informații între rețetele asociate cu persoanele care se află în proximitate una de alta. Adăugarea stratului de pooling le permite rețelelor LSTM să învețe despre interacțiunile dintre oameni și, mai important, despre traiectoriile care se intersectează. Informația din straturile ascunse a rețelelor LSTM învecinate este adunată și împărțită între acestea pentru a obține o evitare fină a coliziunilor. Această metodă a reușit să obțină cele mai mici pierderi în cadrul seturilor de date *ETH* [19] și *UCY* [20], obținând rezultate state-of-the-art.

Asemănător cu [18] este *Social GAN* [21], ce folosește metoda introdusă în [18] ca punct de plecare. Acesta îmbunătățește avantajele aduse de stratul de pooling cu beneficiile aduce de Rețelele Generative Adversariale (GAN). Provocarea este prezicerea traiectoriilor multiple ale aceleiași persoane și determinarea dacă acestea sunt acceptabile din punct de vedere social sau nu. Sistemul este compus din două rețele neurale, un model generativ și unul discriminativ. Modelul generativ se folosește de rețele LSTM pentru fiecare persoană folosind o schemă encoder-decoder ce prezice pozițiile viitoare bazându-se pe observații anterioare. Fiecare LSTM este conectat la un start de pooling pentru a incorpora informații despre alte persoane din scenă. Noutățile aduse stratului de pooling în [21] este că fiecare rețea ia în considerare fiecare persoană din scenă, întrucât oamenii aflați la o distanță mare pot influența traiectoria altei persoane. Informația din stratul de pooling este reprezentată de secvențe de poziții și stări ascunse ce sunt codificate folosind un perceptron multi-strat. Ieșirea rețelei generative este apoi analizată de către rețeaua discriminativă, a cărui scop este de a învăța căi acceptabile din punct de vedere social prin diferențierea dintre traiectoriile generate false și cele reale. Rezultate state-of-the-art au fost obținute în evaluarea sistemului folosind

coordonatele spațiale ale persoanelor din seturile de date ETH [19] și UCY [20]. Metricile de eroare au fost calculate pentru 12 poziții viitoare prezise după observarea a 8 poziții anterioare.

O altă lucrare de cercetare care dorește atenuarea limitărilor lucrării [18] este *Xu et al.* [22]. Aceasta definește o arhitectură care partajează toate informațiile despre persoanele dintr-o scenă. Provocarea este de a modela diferitele nivele de influență pe care oamenii le au asupra unei ținte specifice. Arhitectura constă din 3 module principale: un modul ce codifică mișcarea oamenilor, unul ce modelează afinitatea spațială dintre diferite persoane și unul pentru codificarea mișcării de deplasare. Modulul de codificare folosește rețele LSTM pentru a învăța tipare și de a prezice traiectorii bazându-se pe atribute ale mișcării precum viteza, direcția și accelerația. Modelul de afinitate spațială folosește un perceptron multi-strat pentru a mapa pozițiile persoanelor într-un spațiu de caracteristici cu mai multe dimensiuni. Această transformare este necesară pentru a determina influența pe care o anume persoană o are asupra persoanei țintă (cea a cărei traiectorii este estimată). Modulul de codificare a deplasării combină informația extrasă de celelalte două module și prezice pozițiile viitoare ale unei persoane.

Bartoli at al. [23] îmbunătățește [18] prin adăugarea contextului de mediu. Conceput pornind de la [18], [23] sporește avantajele aduse de stratul de pooling prin adăugarea unui alt strat de pooling pentru obiectele statice din mediu. Predicția traiectoriei persoanelor este obținută prin considerarea atât a pozițiilor, cât și a interacțiunilor anterioare cu obiectele din scenă. Informația este codificată de către rețelele LSTM ce sunt conectate de doua straturi de pooling care codifică două tipuri de interacțiuni: om-om și om-mediu. Stratul care codifică interacțiunile om-mediu modelează nivelul de influență pe care fiecare obiect din mediu îl are asupra persoanei țintă. Obiectele statice care pot influența o traiectorie sunt definite manual în funcție de percepția generală socială și distanțele dintre persoane.

Soluția este creată cu scopul de a funcționa în scenarii aglomerate, predominând imaginile din mediul interior. Evaluarea a analizat 12 poziții prezise după ce s-au observat 8 poziții anterioare. Această metodă depășește rezultate obținute de *Social LSTM* [18] pe un subset din setul de date UCY [20] și un nou set de date ce conține imagini dintr-un muzeu de artă.

În cadrul *Sophie* [24] este propusă o nouă variantă de predicție a traiectoriei ce combină atât informațiile legate de interacțiunile sociale, cât și cele despre constrângerile fizice ale mediului. Această soluție estimează traiectoriile bazandu-se pe imaginile captate de o cameră fixă așezată la înălțime. Arhitectura este compusă din trei module: un extractor de caracteristici, un modul de atenție și un predictor de traiectorie. Extractorul de caracteristici folosește o rețea neurală pentru a determina elementele cheie din imagini, care sunt apoi folosite pentru a codifica pozițiile precedente ale persoanelor și influența pe care o au alte persoane asupra traiectoriei. Modulul de atenție încorporează două componente ce ilustrează influența socială a oamenilor și constrângerile de natură fizică ale scenei, folosind caracteristicile detectate în imagini. Predictorul de traiectorie utilizează o rețea adversarială generativă bazată pe codificările de scurtă și lungă durată pentru a obține mai multe traiectorii posibile pentru o anume persoană. În faza de evaluare, rezultatele obținute de sistemul propus au fost comparate cu cele obținute de *Social GAN* [21] și, deși valorile obținute au fost asemănătoare, *Sophie* s-a descurcat mai bine în medie, obținând erori mai mici pe anumite seturi de date.

Soluțiile din [17], [18], [21], [22], [23] și [24] se limitează doar la imaginile captate de o cameră plasată la înălțime, deoarece această perspectivă oferă posibilitatea pre calcularii coordonatelor spațiale. Atunci când se folosesc în aplicații robotice, pot avea o performanță mai scăzută, imaginile folosite fiind înregistrate de către o cameră de la nivelul omului care este amplasată pe robot. De asemenea,

fluxul de imagini din seturile de date folosite de aceștia sunt obținute folosind o cameră statică, pe când camera robotului poate fi mobilă.

O soluție pentru imaginile cu o perspectivă de la nivelul omului este introdusă în *Yagi et al.* [25]. Provocarea este de a crea o metodă precisă pentru prezicerea traiectoriei bazată pe detecții din video-uri captate din perspectiva platformei robotice. Pentru a obține rezultate conclusive, sistemul augmentează informația legată de poziția persoanelor din imagini cu mai multe date. Pozițiile viitoare a unei anumite persoane sunt estimate prin analiza pozițiilor 2D din imaginile anterioare, a scalei, a posturii și a mișcării camerei. Sistemul a fost testat atât pe seturi de date existente [26] cât și seturi noi de date, spre deosebire de metoda propusă în [18]. Sistemul prezice 10 poziții viitoare ale unei persoane bazate pe 10 poziții anterioare. Soluția a obținut cele mai bune rezultate și cea mai mică eroare comparativ la traiectoria reală în cazul ambelor seturi de date.

3.2 Sisteme de re-identificare a persoanelor

Rețelele neuronale convoluționale sunt folosite cu succes în multe domenii, obținând rezultate bune în multe aplicații practice. Ele sunt de asemenea studiate și în contextul urmăririi și re-identificării de persoane. Lucrările recente prezintă multe variații ale acestora care obțin rezultate bune în cadrul urmăririi și re-identificării de persoane. Următoarea secțiune prezintă cele mai recente tehnici propuse în acest domeniu.

3.2.1 Metode bazate pe extragere de caracteristici

În cadrul *Voigtlaender et. al* [27], un sistem de urmărire prin re-identificare a fost propus. Această metodă este aplicabilă urmăririi unei singure persoane. Arhitectura este construită pornind de la rețeaua Faster R-CNN pre antrenată pentru a detecta 80 de clase de obiecte și care funcționează ca un extractor de caracteristici pentru restul rețelei. Caracteristicile extrase sunt apoi trimise către un modul de detecție în doi pași, compus dintr-un RPN invariant la clasă, urmat de un cap de detecție dependent de categorie. Modulul obține la ieșire chenare candidat pentru imaginea de intrare folosind informații anterioare, care sunt apoi trimise către ultimul modul, Tracklet Dynamic Programming Algorithm (TDPA). Scopul TDPA-ului este de a urmări ținta prin calcularea tracklet-urilor, care sunt traiectorii de scurtă durată. Modulul folosește programarea dinamică pentru a găsi cel mai bun candidat folosind istoricul tracklet-ului. Avantajul metodei propuse este re-detectarea țintei, făcându-l util în cazul în care ținta nu mai poate fi detectată pentru o perioadă mai lungă de timp. De asemenea, prin folosirea tehnicii de programare dinamică, rețeaua este mai rezilientă în cazul obiectelor asemănătoare. Sistemul propus este mai bun decât alte metode state-of-the-art pe zece seturi de date diferite de testare a urmăririi persoanelor.

Yang et al. [28] prezintă un sistem capabil de re-identificarea persoanelor din video-uri. Sistemul propus modelează relațiile temporale dintre cadrele filmului pentru a obține mai multe informații despre persoana urmărită. Legăturile temporale sunt reprezentate folosind Spatial-Temporal Graph Convolutional Network (STGCN). Primul pas al sistemului este extragerea caracteristicilor distincte ale persoanei țintă și gruparea acestora într-o hartă de caracteristici pentru o secvență de imagini folosind o rețea convoluțională precum ResNet50 [29] ca extractor. Componenta principală a arhitecturii este împărțită în trei ramuri: o ramură temporală care extrage indicii temporale din cadrele adiacente, o ramură spațială care extrage caracteristici structurale ale persoanei țintă și o ramură globală care extrage caracteristici de aspect ale acesteia. Atât ramura spațială cât și cea temporală folosesc o rețea convoluțională bazată pe grafuri pentru a modela relațiile.Graful temporal reprezintă legăturile dintre bucăți de imagini din cadre diferite, pe când graful spațial mapează legăturile dintre bucățile de imagine din același cadru. Fiecare imagine din video are un graf spațial asociat. Antrenând rețeaua

calculând erorile pentru fiecare ramură, arhitectura propusă obține rezultate state-of-the-art pe două seturi de date.

Zhong et. al [30] propune o nouă abordare pentru re-identificarea persoanelor prin folosirea unei rețele neuronale numită Align-to-Part Network (APNet). Ideea din spatele APNet este corelarea caracteristicilor regiunilor corpului unei persoane, întrucât pot exista ocluzii care să influențeze aspectul general al acesteia. Chenarele extrase de detectorul de persoane sunt rafinate pentru a acoperi părțile corpului holistice ale oamenilor. Arhitectura este compusă din trei module: un detector de persoane, un aliniator de chenar și un extractor de caracteristici dependent de regiuni. Detectorul de persoane este inspirat de OIM, o rețea care calculează atât chenarele, cât și caracteristicile persoanelor din imagini, antrenat folosind funcția de eroare de la RPN. Aliniatorul de chenar este responsabil de rafinarea chenarelor obținute de la detector, creând chenare pentru fiecare parte vizibilă a corpului. Ultima componentă, extractorul de caracteristici dependent de zonă, extrage caracteristicile vizuale pentru diferitele părți ale corpului pentru o corespondență parțială. Avantajul folosirii acestui modul în loc de extractorul de caracteristici general este supresia extragerii de caracteristici pentru zonele invalide sau ocluzate. Distanța dintre chenarele din diferite cadre reprezintă atât distanța dintre caracteristicile generale cât și distanța dintre caracteristicile părților corpului uman. Experimentele realizate pe două seturi de date arată că se obțin rezultate competitive cu rezultate sistemelor state-of-the-art.

În Gao et. al [31] sarcina de re-identificare a persoanelor este abordată folosind o arhitectură bazată pe mai multe module care combină atât aspectul vizual al persoanelor, cât și caracteristicile poziției corporale. Punctele cheie ale corpului uman sunt extrase folosind metoda OpenPose. Folosind punctele cheie estimate, se calculează caracteristici bazate pe părți. Pentru a depăși problema ocluziilor parțiale, metoda propusă utilizează un modul Pose-Guide Visibility Prediction care prezice care parte a țintei este vizibilă în imagine. Folosind aceste informații, distanța dintre detecțiile din cadre diferite poate fi calculată numai pe baza părților extrase care sunt vizibile simultan în ambele imagini. Componenta de potrivire a caracteristicilor este rezolvată printr-un algoritm de potrivire bazat pe grafuri. Rezultatele obținute de metoda propusă în partea de evaluare sunt comparabilă cu rezultatele de ultimă generație, cu îmbunătățiri asupra problemei de re-identificare a persoanei ocluse.

3.2.2 Metode bazate pe metrici puternice

Wu et. al [32] au propus o abordare bazată pe metrică pentru învățarea nesupravegheată a similarității distanțelor camerelor pe baza imaginilor de intrare. Au introdus o nouă funcție de loss pentru similitudine care poate calcula similitudinea distanțelor *intra-camera* și *cross-camera* pentru potrivirea persoanelor. Metrica propusă ar trebui să atenueze variațiile care provin de la înregistrarea imaginilor pe cameră. Funcția introdusă *Camera-Aware Similarity Consistency Loss* combină trei distanțe de similitudine: consistența similitudinii *intra / cross-camera*, conservarea similitudinii *intra-camere* și funcția de *loss* pentru consistenței similitudinii în funcție de cameră. Pentru a îmbunătăți performanța sistemului, funcția de *loss* este îmbunătățită de o schemă de învățare *coarse-to-fine*. Rolul schemei de învățare este de a obține distribuții consistente de similaritate între perechi de imagini. Sistemul obține rezultate state-of-the-art pe două seturi de date dificile.

Zeng et. al [33] au propus o metodă de învățare nesupravegheată care utilizează clustering ierarhic pentru a re-identifica oamenii. Clusterizarea se face folosind caracteristici extrase de o rețea ResNet-50 [29]. Modelul utilizează un algoritm în buclă care include un mecanism de eșantionare PK pentru a genera un nou set de date de antrenare pentru metoda de clustering la fiecare iterație. Funcția de *loss* utilizată în antrenarea metodei propuse este funcția de *loss hard-batch triplet loss*, care, în combinație cu eșantionarea PK, poate discrimina exemple dificile, îmbunătățind în același timp clusterizarea. Mai

mult, pentru a reduce riscul de pseudo-etichete false, toate etichetele sunt inițializate la începutul fiecărei iterații a algoritmului propus. Pentru faza de evaluare, au fost utilizate 16 clase pentru mecanismul ierarhic de grupare. Sistemul a obținut rezultate state-of-the-art comparativ cu alte sisteme de re-identificare cu învățare nesupravegheată.

3.2.3 Metode alternative

Bergenmann et. al [34] a introdus o nouă metodă ce nu folosește un algoritm de urmărire dedicat și, implicit, nu are nevoie de o fază de antrenare dedicată pentru problema urmăririi și re-identificării persoanelor. Abordarea lor demonstrează că traiectoriile obiectelor pot fi obținute folosind regresorul unui detector de obiecte, în cazul acesta prin folosirea regresorului de la Faster R-CNN. Presupunerea pe care autorii acestei lucrări o fac este că persoanele nu se deplasează foarte mult de la un cadru la altul. Deoarece diferența dintre pozițiile unei persoane în cadre consecutive este mică, un regresor este suficient pentru a putea re-identifica aceeași persoană. Pentru a combate efectele frecvenței mici cu care camerele înregistrează, autorii actualizează viteza tuturor obiectelor din imagine. De asemenea, pentru a îmbunătății mai mult sistemul, autorii au introdus un model bazat pe rețele Siamese pentru re-identificarea de scurtă durată. Rezultatele obținute de acest sistem sunt mai bune decât a metodelor state-of-the-art pe scenarii de urmărire dificile.

Braso et. al [35] se concentrează pe definirea unui framework complet diferențiabil bazat pe Message Passing Networks (MPNs) pentru urmărirea mai multor obiecte. Metoda propusă folosește un graf pentru a mapa relațiile dintre persoanele detectate în cadre diferite pentru a îi putea re-identifica. Nodurile din graf constau în caracteristicile extrase de o rețea convoluțională din chenarele extrase de detectorul de persoane. Graful inițial are o structură complet conectată, cu laturi între fiecare pereche de chenare. Folosind o tehnică de transmitere de mesaje printr-o rețea neuronală, structura grafului este actualizată la arhitectura corectă. În cadrul acestui pas, nodurile partajează informații vizuale precum caracteristici de aspect pe lângă informații geometrice despre laturi. Autorii propun o ajustare a rețelelor de transmitere de mesaje prin alterarea pasului de actualizare a nodului folosind informații de timp. La sfârșitul tehnicii de transmitere de mesaje ce ține cont de timp, laturile rămase din graf reprezintă corelarea dintre persoanele din cadre diferite. Experimentele conduse arată că această metodă este mult mai bună decât alte tehnici state-of-the-art pe trei seturi de date din punct de vedere al metricilor MOTA și IDF1.

O abordare similară a problemei de urmărire și re-identificare de persoane este propusă în [36]. Scopul acestei lucrări este de a crea un sistem ce învață relațiile și topologia punctelor cheie pentru o fi mai robust în cazul ocluziei persoanelor urmărite din imagini. Sistemul folosește o arhitectura multi-modulară compusă din trei componente principale: un modul semantic al cărui scop este extragerea de caracteristici, un modul relațional responsabil de modelarea relațiilor dintre caracteristicile locale, și un modul topologic ce se ocupă de alinierea și predicția similitudinii dintre imagini. Metoda folosește ResNet-50 [29], o rețea neuronală convoluțională, pentru a extrage caracteristicile persoanelor detectate și un estimator al punctelor cheie, HR-Net [37] mai precis, pentru estimarea caracteristicilor locale. Pentru a învăța relațiile dintre caracteristicile locale, autorii propun folosirea Adaptive Directed Graph Convolution Layer, al cărui scop este adaptarea mecanismului de transmitere de mesaje în așa fel încât caracteristicile semantice să fie mai importante decât caracteristicile ce nu aduc vreun beneficiu. Învățarea topologiei umane este realizată prin folosirea unui nou strat cross-graph embedded-alignment, care calculează caracteristicile aliniate la topologia umană prin folosirea caracteristicilor sub-grafice ale imaginii de intrare. Rezultatele experimentale obținute de sistemul propus depășesc rezultatele metodelor state-of-the-art pe un set de date cu multe ocluzii.

3.2.4 Comparația sistemelor identificate

Rețelele neurale convoluționale oferă soluții robuste pentru sistemele de urmărire și re-identificare a persoanelor. Multe dintre soluțiile prezentate recent utilizează rețele convoluționale adânci pentru a extrage cele mai relevante caracteristici pentru persoanele detectate astfel încât să obțină cele mai bune rezultate la faza de potrivire a caracteristicilor. Mai mult, cercetătorii folosesc caracteristici parțiale, corespunzătoare punctelor cheie pentru potrivirea separată a caracteristicilor pentru a trata ocluziile și variația vizuală. Este evident că rezultatele obținute de aceste sisteme sunt mult mai exacte și mai robuste. Comparația dintre sistemele recente de urmărire și re-identificare a persoanelor este prezentată în tabelul 3.1.

Sistem	**An**	**Acuratețe**	**Observații**
Siam R-CNN [27]	2020	63.9% AUC pe NfS, 64.9% pe UAV123, 70.1% pe OTB2015, 62.4% pe VOT2018	
STGCN [28]	2020	83.70% mAP pe MARS, 95.70% pe DukeMTMC-VideoReID	Numărul de straturi și zone din GCN este o variabilă fixă ce poate influența rezultatele
APNet [30]	2020	88.9% mAP pe CUHK-SYSU, 41.9% pe PRW	
PVPM [31]	2020	61.2% mAP pe Occluded-REID, 72.3% pe Partial-REID, 29.2% pe P-DukeMTMC-reID	Numărul de părți ales are un impact mare asupra rezultatului
Camera-Aware Similarity Consistency [32]	2019	35.5% mAP pe Market-1501, 37.8% pe DukeMTMC	Poate fi îmbunătățit printr-o învățare de consistență în doi pași coarse-to-fine
HCT [33]	2020	56.4% mAP pe Market-1501, 50.7% mAP pe DukeMTMC-reID	Are nevoie de numărul corect de clustere
Tracktor [34]	2019	44.1% MOTA pe 2D MOT2015, 54.4% pe MOT16, 53.5% pe MOT17	Persoana trebuie să se miște foarte puțin între consecutive imagini
Neural Solver for MOT [35]	2020	51.5% MOTA pe 2D MOT2015, 58.6% pe MOT16, 58.8% pe MOT17	
High-Order Information Matters [36]	2020	84.9% mAP pe Market-1501, 75.6% pe DukeMTMC, 91.0% pe Partial-REID, 86.4% pe Partial-iLIDS	

Tabel 3.1 Comparație între sistemele de urmărire pe baza de re-identificare a persoanelor.

4. Arhitectura propusă pentru urmărirea persoanelor

Pentru a rezolva problema detecției și urmăririi de persoane propunem două arhitecturi, una bazată pe predicția traiectoriei, și una bazată pe re-identificare.

4.1 Arhitectură pe baza predicției traiectoriei

Figura 4.1 prezintă arhitectura propusă pentru a rezolva problema urmăririi de persoane pe baza predicției traiectoriei.

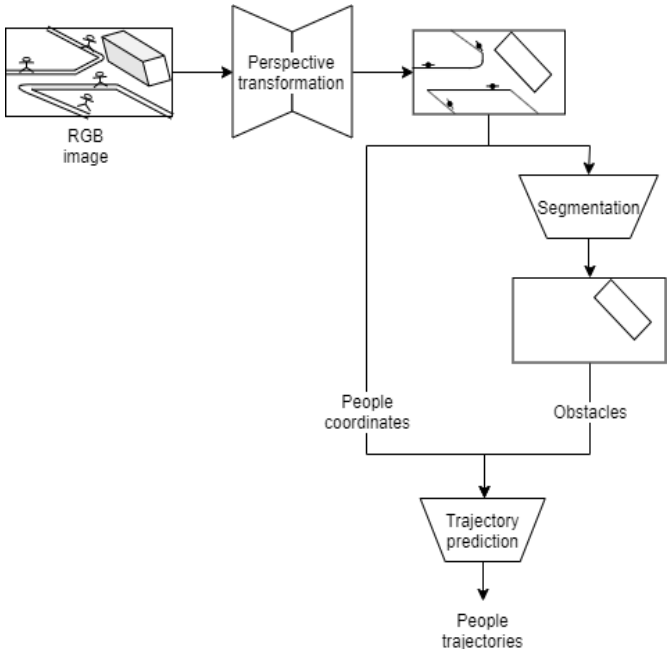

Figura 4.1 Arhitectura propusă pe baza predicției traiectoriei.

Arhitectura preia ca intrare un flux de imagini RGB și procesează un cadru la un moment dat. Prima fază a metodei propuse este o transformare a perspectivei, care transformă un cadru preluat dintr-un unghi de la nivelul unui om într-o imagine văzută de sus. Imaginea rezultată are aceleași caracteristici și distanțe relative între oameni și obiectele existente ca și cea inițială. Imaginea procesată este apoi trecută printr-o rețea de segmentare pentru a extrage obstacolele din scenă. Segmentarea va utiliza un model similar unei rețele complet convoluționale și va genera o imagine care conține doar obstacolele din scenă. În paralel cu operația de segmentare, coordonatele persoanelor din imagine vor fi extrase ca poziții 2D. Ultima fază este sistemul pentru predicția traiectoriei. Pentru a prezice traiectorii pe baza coordonatelor extrase, sistemul va utiliza două straturi de pooling suplimentare, unul pentru interacțiunile sociale și unul pentru obstacole. Rețeaua va fi similară cu cea din Social GAN [22], cu o ajustare pentru a lua în considerare obstacolele din scenă.

4.2 Arhitectură pe baza re-identificării persoanelor

Figura 4.2 prezintă arhitectura propusă pentru a rezolva problema urmăririi de persoane pe baza re-identificării persoanelor. Arhitectura preia imagini RGB care sunt trecute prin două module, unul de extragere a joint-urilor persoanelor și unul de segmentare a imaginilor. Imaginile segmentate furnizează informațiile necesare pentru a stabili care joint-uri sunt vizibile în imagine și care sunt ascunse de obstacole. După identificarea joint-urilor vizibile ale persoanelor vor fi extrase caracteristici importante din imagini în acele zone identificate de joint-uri. Caracteristicile vor fi trecute printr-un

modul de matching care va ține cont și de componenta temporală a acestora. În urma potrivirii caracteristicilor vor fi re-identificate persoanele în imagini consecutive.

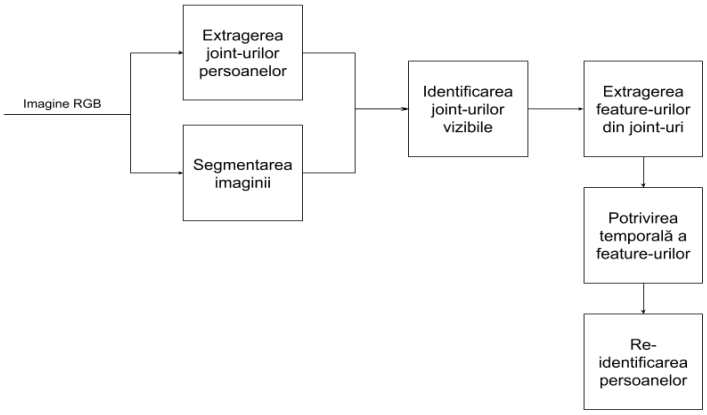

Figura 4.2 Arhitectura propusă pe baza re-identificării de persoane.

5. Seturi de date

5.1 Seturi de date pentru detecția pietonilor

Există mai multe tipuri de seturi de date. Pentru recunoaștere de imagini, cel mai folosit set este ImageNet [38] . Pentru detecție de obiecte generice, Pascal VOC [39] , COCO [40] , SUN [41] sunt seturile de date cele mai folosite, conțin zeci de mii de imagini și obiecte și sunt folosite pentru diverse obiecte , de la persoane și oameni până la animale, biciclete, inclusiv obiecte de uz caznic. Pe de altă parte, KITTI [42] , CityScapes [43] , BDD100K [44] sunt seturi de date care se folosesc în special pentru scenarii legate de mașini autonome. Aceste seturi de date conțin doar imagini filmate în trafic. Există și seturi de date specializate pentru ceea ce ne dorim noi, detecție de persoane – Eurocity [45] (50000 de imagini) , INRIA [46] (600 de imagini) , Caltech Pedestrian Dataset [47] (250000 de imagini obținute din filmări) și ETH [48] (2500 de imagini). Acestea au imagini adnotate doar cu persoane. Există și seturi de date folosite pentru urmărirea persoanelor, de exemplu MOT (Multiple object tracking). [49] În cadrul proiectului, ne dorim să folosim seturi de date deja existente, cum ar fi cele menționate anterior, dar și seturi de date existente sau în curs de dezvoltare înregistrate în cadrul campusului nostru universitar.

5.2 Seturi de date pentru urmărirea persoanelor

Pentru a putea proiecta o arhitectură corespunzătoare obiectivului de urmărire de persoane, un pas prealabil de căutare și analiză a seturilor de date existente a fost necesar. Deoarece scopul acestei lucrări este de a crea un sistem capabil de urmărirea de persoane prin folosirea informațiilor vizuale, seturile de date explorate conțin imagini cu oameni atât în mediul interior, cât și în cel exterior.

Cele mai importante caracteristici urmărite ale seturilor de date sunt unghiul vizual al camerei și mobilitatea acesteia (fixă sau mobilă). Din punct de vedere al unghiului vizual, majoritatea secvențelor de imagini au fost înregistrate fie dintr-o perspectivă de la înălțime, fie dintr-o perspectivă de la nivelul omului. Din punct de vedere al mobilității, există mai multe seturi de date ce folosesc o cameră statică decât cele care folosesc o cameră mobilă.

5.2.1 Seturi de date cu o perspectivă de la înălțime

Setul de date ETH - BIWI Walking Pedestrians [19] conține imagini colectate din două scenarii diferite din mediul exterior, așa cum se poate observa în figura 5.1. Setul de date conține 32.324 imagini cu rezoluția de 640x480 pixeli. Acesta prezintă scenarii cu o densitate mare a oamenilor și obiecte mici și

răsfirate (copaci, stâlpi, bănci). Adnotările conțin atât informații despre poziția și viteza oamenilor din imagini, cât și informații despre grupuri (oameni care par că merg împreună). De asemenea, persoanele detectate au asociat un număr de identificare.

Figura 5.1. Imagini extrase din setul de date *ETH - BIWI Walking Pedestrians.*

Setul de date UCY - Crowds by example [20] include imagini din 3 scenarii diferite, așa cum este prezentat în figura 5.2. Acesta este alcătuit din 25.853 imagini cu rezoluția de 576x720 pixeli, grupate în 4 secvențe video. Imaginile includ aproximativ 10-15 persoane simultan în scenarii din mediul extern, cu puține obstacole. Adnotările includ puncte 2D ale persoanelor și ID-urile corespunzătoare acestora.

Figura 5.2. Imagini extrase din setul de date *UCY - Crowds by example.*

Setul de date Grand Central [50] este o colecție de imagini obținute în interiorul unei gări mari din New York, traversată de sute de persoane, fapt vizibil în figura 5.3. Acesta include 50.010 imagini cu o rezoluție de 720x480 pixeli. Camera a fost amplasată la înălțime, iar zona captată conține un obstacol mare în centrul imaginii. Scena este foarte aglomerată, încadrând zeci de oameni într-un singur cadru.

Figura 5.3. Imagini extrase din *Grand Central.*

Alt set de date înregistrat folosind o cameră fixă plasată la înălțime este reprezentat de MuseumVisitors [51], exemple din acesta fiind în figura 5.4. MuseumVisitors este un set de date mic, acesta conținând doar 4.820 imagini cu o rezoluție de 1280x800 pixeli. Imaginile au fost captate într-un mediu interior, într-un muzeu, cadrul conținând câteva obstacole reprezentate de către exponate și puține persoane detectate (între 5 și 10 în medie). Adnotările conțin chenarele încadratoare ale persoanelor din cadru, numărul lor de identificare și informații suplimentare despre apartenența lor la un grup de oameni.

Figura 5.4. Imagini extrase din *MuseumVisitors.*

5.2.2 Seturi de date cu o perspectivă de la nivelul omului

PETS2009 [52] este un set de date mic ce conține un mix de imagini de la mai multe camere ce înregistrează simultan, atât de la nivelul omului, cât și de la înălțime. Pozițiile camerelor se pot observa în figura 5.5. Setul de date este compus dintr-o colecție de imagini cu o rezoluție de 720x576 și 768x576 pixel, rezoluția depinzând de cameră, ce au fost captate într-un mediu exterior. 2.857 din imagini au fost înregistrate de camerele de la înălțime, iar restul de 795 de cele de la nivelul omului. Sunt prezente câteva obstacole în cadre, reprezentate de copaci și stâlpi.

Figura 5.5. Imagini extrase din setul de date *PETS2009.*

MOT17: Multi-Object Tracking [53] este un set de date mare și dificil, ce conține secvențe cu perspective mixte (exemple în figura 5.6). Setul de date conține 11235 imagini în total, împărțite în secvențe ce captează diferite scenarii. Fiecare scenă are o configurație diferită, întrucât setul de date este alcătuit din imagini ce au fost obținute atât în mediul interior, cât și în cel exterior. Numărul de persoane simultan în cadru variază în funcție de scenă. De asemenea, camerele folosite sunt atât statice, cât și mobile, cele mobile find mișcate de către o platformă robotică. Mare parte din imaginile din acest set de date au o rezoluție de 1920x1080 pixeli, excepție făcând o secvență cu rezoluția de 640x480 pixeli. Adnotările conțin informații despre pozițiile 2D ale persoanelor din imagini și numărul de identificare corespunzător.

Figura 5.6. Imagini extrase din setul de date *MOT17: Multi-Object Tracking.*

Un set de date ce conține doar imagini captate de la nivelul omului este ETH - Social behaviour modeling [54]. Acesta conține imagini din mediul exterior în multiple locații, ca în figura 5.7. Este un set de date relativ mic, conținând doar 5.361 imagini la o rezoluție de 640x480 pixeli. Sunt prezente obstacole rare, precum bănci și copaci. Persoanele sunt reprezentate în adnotări prin poziția lor în imagine.

Figura 5.7. Imagini extrase din setul de date *ETH - Social behaviour modeling.*

În setul de date First-Person Social Interactions [26] imaginile au fost obținute într-un parc de distracții cu un număr semnificativ de oameni. Totuși, imaginile pot conține atât un număr mic de oameni în cadru sau, dimpotrivă, un număr foarte mare, acest lucru fiind dependent de locul din parc în care s-a făcut filmarea. Figura 5.8 prezintă exemple de imagini din acest set de date. Setul de date conține imagini dificile, întrucât au fost înregistrate în timp ce se aștepta la coadă, sau când se călătorea cu trenul. Setul de date conține atât scene din mediul interior, cât și din cel exterior, iar acestea au fost înregistrate exclusiv de la nivelul omului.

Figura 5.8. Imagini extrase din setul de date *First-Person Social Interactions.*

5.2.3 Seturi de date de la nivelul omului și informații de adâncime

JRDB [55] este un set de date mare, conținând aproximativ 55.000 imagini. Acesta a fost obținut folosindu-se o platformă robotică cu o multitudine de camere și senzori, astfel încât setul conține mai multă informație decât cea de culoare RGB. Un exemplu de imagine din acest set de date se poate vedea în figura 5.9. Imaginile au fost captate simultan folosind 5 camere, dar doar o singură cameră a captat și informații de adâncime. Setul de date conține atât imagini din mediul interior, cât și din cel exterior, iar numărul de persoane ce apar simultan în cadru variază între 0 și 30. Imaginile pot include și obstacole precum scaune sau mese, dar acest lucru depinde de scenă. Adnotările conțin informații despre pozițiile 3D ale persoanelor din imagine și numărul de identificare corespunzător.

Figura 5.9. Imagine extrasa din setul de date *JRDB.*

Un alt set de date ce augmentează datele RGB cu informații de adâncime este IAS-Lab People Tracking [56]. Acesta conține imagini captate de o cameră fixă într-un mediu interior, mai precis într-un laborator, cu un număr mic de oameni, precum se poate vedea în figura 5.10. Setul de date combină 4.671 imagini din 3 scenarii diferite: traiectorii simple, interacțiuni între persoane, obstacole și interacțiuni între persoane. Rezoluția imaginilor RGB este de 640x480 pixeli, iar calitatea informațiilor de adâncime este de 160x120 pixeli. Adnotările conțin pozițiile oamenilor din imagine și ID-ul acestora.

Figura 5.10. Imagine extrasa din setul de date *IAS-Lab People Tracking.*

5.2.4 Setul de date colectat în cadrul proiectului

În cadrul proiectului am colectat un set de date pentru conducerea autonomă, care va fi folosit în lucrările viitoare pentru experimentele noastre. Pentru aceste experimente, am folosit un videoclip din campusul nostru format din 13001 cadre, care vor fi menționate în continuare ca set de date POLI. Imaginile conțin în principal mașini și pietoni, dar și câteva biciclete și indicatoare, un număr total de 60227 obiecte, din care 41064 mașini și 14576 persoane. Pentru generarea ground truth, am adnotat manual fișierul imagini folosind CVAT, un instrument pentru generarea ground truth. În setul nostru de date, peste 90% dintre obiecte sunt fie mașină, fie persoană.

În tabelul 5.1 avem toate rezultatele pentru setul de date POLI, inclusiv repere cu doar cele două etichete - auto și persoană. Pentru Yolo, MaP de 0,55 este foarte aproape de 0,54, media MaP-urilor pentru setul de date BDD100K. Însă MaR este aproape dublu față de valoarea - un motiv pentru această fiind faptul că există mai puține obiecte în setul nostru de date și cele mai multe dintre ele sunt mari și aproape de mașină. Numerele sunt mai mari și pentru 2 clase, rechemarea fiind de 0,40 în loc de media MaRs de 0,36.

	AP@0.50IOU	MAP	AR@0.50IOU	MAR
Yolo	0.69	0.55	0.59	0.47
SSD	0.79	0.65	0.12	0.10
Faster R-CNN	0.68	0.54	0.16	0.13
RetinaNet	0.17	0.26	0.06	0.09
Yolo (car and person)	0.71	0.57	0.62	0.50
SSD (car and person)	0.90	0.74	0.13	0.10
Faster R-CNN (car and person)	0.80	0.63	0.17	0.13
RetinaNet (car and person)	0.20	0.30	0.06	0.10

Tabelul 5.1 Evaluarea setului de date propriu

6. Concluzii

În cadrul primei etape a proiectului s-au analizat diferite arhitecturi de rețele neurale adânci pentru detecția de persoane și pentru urmărirea de persoane, atât pentru spații închise (indoor) – caz potrivit pentru aplicații ale roboților sociali, cât și spații deschise (outdoor) – caz potrivit pentru pietoni în aplicații de conducere autonomă. S-au comparat și evaluat aceste arhitecturi și s-au propus arhitecturi pentru detecția și urmărirea persoanelor.

S-au investigat diferite seturi de date existente în domeniul public, s-au comparat aceste seturi de date din punct de vedere al performanțelor realizate pe diferite arhitecturi adânci și s-a colectat propriul set de date pentru detecția pietonilor și a mașinilor. S-a evaluat setul de date colectat.

În Etapa 2 (2021) vom trece la implementarea soluțiilor de detecție și urmărire a persoanelor indoor și outdoor.

Bibliografie

1. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
2. J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. IJCV, 2013.
3. R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015.
4. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
5. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.
6. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016.
7. Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. CoRR, abs/1804.02767, 2018.
8. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," arXiv:1512.02325, 2015.
9. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. ´ Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017.
10. T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and ´ S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017.
11. Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In ECCV, 2018.
12. Hei Law, Yun Teng, Olga Russakovsky, and Jia Deng. Cornernet-lite: Efficient keypoint based object detection. CoRR, 2019.
13. X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019.
14. D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. In ECCV 2016, pages 749–765. Springer, 2016.
15. G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking," in 2017 IEEE International Symposium on Circuits and Systems (ISCAS), May 2017.
16. Wojke, N., Bewley, A., Paulus, D.: 'Simple online and realtime tracking with a deep association metric'. Proc. Int. Conf. on Image Processing, Beijing, China, 2017.
17. T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Tracking by prediction: A deep generative model for multi-person localisation and tracking," in2018 IEEE WinterConference on Applications of Computer Vision (WACV), pp. 1122–1132, March 2018.
18. A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in2016 IEEE Conference onComputer Vision and Pattern Recognition (CVPR), pp. 961–971, June 2016.21.
19. S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in2009 IEEE 12th International Conference on Computer Vision, pp. 261–268, Sep. 2009.
20. A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example,"Computer graphics forum, vol. 26, pp. 655–664, Sep. 2007.
21. A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in2018 IEEE/CVFConference on Computer Vision and Pattern Recognition, pp. 2255–2264, June 2018.
22. Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5275–5284, June 2018.
23. F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction,"in2018 24th International Conference on Pattern Recognition (ICPR), pp. 1941–1946,Aug. 2018.
24. A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, "Sophie: An attentive GAN for predicting paths compliant to social and physical constraints,"CoRR,vol. abs/1806.01482, June 2018.
25. T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in2018 IEEE/CVF Conference on Computer Vision and PatternRecognition, pp. 7593–7602, June 2018.

26. A. Fathi, J. K. Hodgins, and J. M. Rehg, “Social interactions: A first-person perspective,”in2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1226–1233,June 2012.20.
27. P. Voigtlaender, J. Luiten, P. Torr, and B. Leibe, “Siam r-cnn: Visual tracking byre-detection,” pp. 6577–6587, 06 2020.
28. J. Yang, W.-S. Zheng, Q. Yang, Y. Chen, and Q. Tian, “Spatial-temporal graph convolutional network for video-based person re-identification,” pp. 3286–3296, 06 2020.
29. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,”in2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pp. 770–778, 2016.
30. Y. Zhong, X. Wang, and S. Zhang, “Robust partial matching for person search in the wild,” pp. 6826–6834, 06 2020.
31. S. Gao, J. Wang, H. Lu, and Z. Liu, “Pose-guided visible part matching for occluded person reid,” in2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 11741–11749, 2020.
32. A. Wu, W. Zheng, and J. Lai, “Unsupervised person re-identification by camera-aware similarity consistency learning,” in2019 IEEE/CVF International Conference onComputer Vision (ICCV), pp. 6921–6930, 2019.
33. K. Zeng, M. Ning, Y. Wang, and Y. Guo, “Hierarchical clustering with hard-batch triplet loss for person re-identification,” pp. 13654–13662, 06 2020.
34. P. Bergmann, T. Meinhardt, and L. Leal-Taix´e, “Tracking without bells and whistles,”CoRR, vol. abs/1903.05625, 2019.
35. G. Bras´o and L. Leal-Taix´e, “Learning a neural solver for multiple object tracking,”pp. 6246–6256, 06 2020.
36. G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun,“High-order information matters: Learning relation and topology for occluded person re-identification,” pp. 6448–6457, 06 2020.
37. S. ke, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” pp. 5686–5696, 06 2019.
38. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. CVPR. 2009.
39. Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision (IJCV), 88(2):303–338, 2010.
40. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. ECCV. 2014.
41. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo”. CVPR. 2010.
42. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. IJRR. 2013.
43. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. CVPR. 2016.
44. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T. BDD100K: A diverse driving video database with scalable annotation tooling. CoRR. 2018.
45. Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M. Gavrila. The eurocity persons dataset: A novel benchmark for object detection. CoRR, abs/1805.07193. 2018.
46. N. Dalal and B. Triggs, Histograms of oriented gradients for human detection. CVPR. 2005.
47. P. Dollar, C. Wojek, B. Schiele, and P. Perona, Pedestrian detection: A benchmark. CVPR. 2009.
48. Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, pages 1–8, 2007.
49. A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.
50. B. Zhou, X. Wang, and X. Tang, “Understanding collective crowd behaviors: Learninga mixture model of dynamic pedestrian-agents,” in2012 IEEE Conference on ComputerVision and Pattern Recognition, pp. 2871–2878, June 2012.
51. F. Bartoli, G. Lisanti, L. Seidenari, S. Karaman, and A. Del Bimbo, “Museum visitors: A dataset for pedestrian and group detection, gaze estimation and behavior understanding,”in2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), pp. 19–27, June 2015.

52. J. Ferryman and A. Shahrokni, “Pets2009: Dataset and challenge,” in2009 TwelfthIEEE International Workshop on Performance Evaluation of Tracking and Surveillance,pp. 1–6, Dec. 2009.
53. A. Milan, L. Leal-Taix´e, I. Reid, and S. Roth, “Mot16: A benchmark for multi-object tracking,”CoRR, vol. abs/1603.00831, March 2016.
54. A. Ess, B. Leibe, K. Schindler, and L. Van Gool, “A mobile vision system for robust multi-person tracking,” in2008 IEEE Conference on Computer Vision and PatternRecognition, pp. 1–8, June 2008.
55. R. Mart´ın-Mart´ın, H. Rezatofighi, A. Shenoi, M. Patel, J. Gwak, N. Dass, A. Federman,P. Goebel, and S. Savarese, “Jrdb: A dataset and benchmark for visual perception for navigation in human environments,”CoRR, vol. abs/1910.11792, Oct. 2019.
56. M. Munaro, F. Basso, and E. Menegatti, “Tracking people within groups with rgb-d data,” in2012 IEEE/RSJ International Conference on Intelligent Robots and Systems,pp. 2101–2107, Oct. 2012.