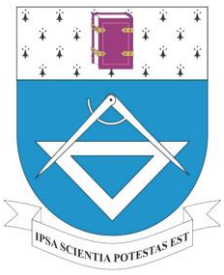

UNIVERSITATEA TEHNICĂ "GHEORGHE ASACHI"
DIN IAȘI

EARLY FUSION BASED DEEP LEARNING TECHNIQUES FOR COMPLEX APPLICATIONS

(TEHNICI DE ÎNVĂȚARE PROFUNDĂ BAZATE PE FUZIUNEA DATELOR LA INTRARE PENTRU APLICAȚII COMPLEXE)

REZUMAT

Adrian-Paul BOTEZATU

Conducător de doctorat: Prof. dr. ing. Adrian BURLACU

IAŞI, 2026

EARLY FUSION BASED DEEP LEARNING TECHNIQUES FOR COMPLEX APPLICATIONS

TEHNICI DE ÎNVĂȚARE PROFUNDĂ BAZATE PE FUZIUNEA DATELOR LA INTRARE PENTRU APLICAȚII COMPLEXE (in Romanian)

Adrian-Paul BOTEZATU
Domeniul: Ingineria Sistemelor

Președinte comisie doctorat: Prof. dr. ing. Vasile-Ion Manta
Universitatea Tehnică „Gheorghe Asachi” Iași

Conducător de doctorat: Prof. dr. ing. Adrian Burlacu
Universitatea Tehnică „Gheorghe Asachi” Iași

Referenți oficiali: Prof. dr. ing. Mihnea-Alexandru Moisescu
Universitatea Națională de Știință și Tehnologie ,,Politehnica” București

Prof. dr. ing. Levente Tamás
Universitatea Tehnică din Cluj-Napoca

Conf. dr. ing. Lavinia-Eugenia Ferariu
Universitatea Tehnică „Gheorghe Asachi” Iași

Comisia de îndrumare și integritate academică:

Prof. dr. ing. Constantin-Florin Căruntu
Universitatea Tehnică „Gheorghe Asachi” Iași

Prof. dr. ing. Vlad Mureșan
Universitatea Tehnică din Cluj-Napoca

Conf. dr. ing. Carlos-Mihai Pascal
Universitatea Tehnică „Gheorghe Asachi” Iași

Cuprins

Capitolul 1

Introducere

Motivație

Inteligența artificială (IA) a schimbat modul în care sistemele inteligente interacționează cu mediul, în special prin metode de învățare profundă (Deep Learning), care permit extragerea automată de caracteristici din seturi mari de date. Rețelele Neurale Convoluționale (CNN) au condus la performanțe superioare în aplicații precum diagnosticarea medicală [1], conducerea autonomă [2], manipularea robotică [3] și servoing-ul vizual [4, 5].

În robotică, învățarea profundă a condus la apariția de sisteme mai precise și adaptabile. Servoing-ul vizual, care utilizează feedback-ul vizual pentru controlul mișcării robotice, beneficiază direct de aceste progrese, în special în medii dinamice sau parțial ocluzate [6]. Similar, în vehiculele autonome, clasificarea suprafețelor rutiere este vitală pentru siguranță și adaptabilitate, însă metodele tradiționale bazate exclusiv pe imagini RGB sunt limitate în condiții dificile. CNN-urile oferă o robustețe și o generalizare superioară, evidențiind potențialul acestora pentru astfel de aplicații [7, 8].

Această teză investighează două aplicații esențiale, servoing vizual și clasificarea suprafețelor rutiere, din perspectiva învățării profunde, evidențiind evoluția academică și industrială a acestor direcții.

Descrierea problemei și contribuții principale

Metodele tradiționale întâmpină limitări semnificative în medii reale, afectate de ocluziuni, iluminare variabilă sau condiții atmosferice. Această teză propune noi arhitecturi CNN pentru a crește robustețea percepției vizuale atât în robotică, cât și în sistemele automotive.

Contribuțiile principale sunt:

- **Dezvoltarea unei arhitecturi CNN cu fuziune timpurie**: Se propune o arhitectură CNN nouă care utilizează o strategie de fuziune a intrare a datelor. Această abordare integrează date vizuale direct în primul strat neuronal al rețelei.

- **Analiză cuprinzătoare a caracteristicilor vizuale suplimentare**: Se oferă o analiză sistematică și evaluare a diferitelor caracteristici vizuale care pot servi ca date suplimentare în arhitectura cu fuziune la intrare a datelor. Aceasta include o comparație detaliată a hărților de nuanțe de gri, hărților de gradient, hărților de segmentare, hărților de puncte de interes și hărților de momente de imagine, examinând efectele lor individuale și combinate asupra acurateței percepției și controlului în diverse scenarii.

- **Revizuire amplă a progreselor în învățarea profundă pentru analiza suprafețelor de drum**: Se realizează o analiză detaliată și o evaluare critică a tehnicilor de învățare profundă pentru clasificarea suprafețelor de drum, identificând principalele provocări și oportunități de îmbunătățire.

- **Implementarea în timp real pentru sistemele de fuziune la intrare a datelor**: Se dezvoltă o implementare a tehnicilor de fuziune la intrare a datelor în sistemele de control robotic în timp real. Această contribuție include strategii de integrare hardware, tehnici de optimizare computațională și proiecte de arhitectură software care permit implementarea practică a abordărilor teoretice propuse pe platforme robotice fizice.

- **Metodologie de generare a datelor sintetice pentru antrenarea servoing-ului vizual**: Se dezvoltă o abordare nouă pentru generarea de date sintetice de antrenament prin transformări homografice, abordând provocarea critică de a obține mostre de antrenament diverse și precise pentru sistemele de servoing vizual. Această metodologie permite antrenarea eficientă a modelelor de învățare profundă fără a necesita colectarea manuală extinsă de date, reducând semnificativ timpul de dezvoltare și îmbunătățind capacitățile de generalizare.

- **Arhitectură hibridă de servoing vizual integrând SuperPoint**: Se propune o arhitectură hibridă pentru sistemele de servoing vizual prin incorporarea SuperPoint, o rețea neuronală auto-supervizată , pentru detectarea și descrierea robustă a punctelor de interes. Această integrare îmbunătățește semnificativ robustețea și precizia metodelor de servoing vizual, permițând sistemelor robotice să facă față mai bine mediilor dinamice.

Structura tezei și rezumatul publicațiilor

Teza este organizată în șapte capitole care acoperă aspectele teoretice și experimentale ale arhitecturilor propuse:

Capitolul 1 introduce motivația integrării Învățării profunde în sistemele inteligente pentru robotică și conducere autonomă, subliniind provocările din servoing-ul vizual și clasificarea suprafețelor rutiere, precum și soluțiile propuse în această lucrare.

Capitolul 2 oferă o bază teoretică solidă privind învățarea profundă, concentrându-se pe rețele neuronale convoluționale și arhitecturi relevante precum AlexNet și ResNet-50. De asemenea, sunt discutate aplicații ale acestor tipuri de rețele în servoing-ul vizual și tendințele recente din literatură. **Capitolul 3** oferă o analiză a problemelor deschide în cele două domenii studiate în această teză de doctorat

Capitolul 4 prezintă arhitectura CNN cu fuziune timpurie propusă, demonstrând beneficiile integrării canalelor vizuale suplimentare la nivelul de intrare (niveluri de gri, gradient) pentru clasificarea suprafețelor rutiere. Evaluările experimentale se bazează pe următoarele lucrări:

- **P. Botezatu**, A. Burlacu, and C. Orhei, "A review of deep learning advancements in road analysis for autonomous driving," Applied Sciences, vol. 14, no. 11, p. 4705, 2024.

- **A.-P. Botezatu** and A. Burlacu, "CNN-Based Early Fusion with Intensity Features for Road Surface Classification" (trimisă la ICSTCC 2025)

Capitolul 4 investighează paradigmele de servoing vizual (IBVS, PBVS, hibride), evidențiind limitările curente și oportunitățile de integrare a rețelelor neuronale convoluționale. Conținutul se bazează pe:

- **P. Botezatu** and A. Burlacu, "A short review of deep learning methods in visual servoing systems," Bulletin of the Polytechnic Institute of Iași, vol. 69, no. 3, pp. 113–136, 2023.

- F.-A. Brașoveanu, A.-I. Iancu, **A.-P. Botezatu**, and A. Burlacu, "3-d vision-based workspace with deep learning capabilities for autonomous robot manipulation," Springer Nature, 2025.

Capitolul 5 detaliază implementarea unui sistem de servoing vizual cu arhitectură CNN și fuziune timpurie, incluzând hărți de regiuni, momente și puncte caracteristice, utilizând o strategie de antrenare în două etape. Rezultatele provin din:

- **P. Botezatu**, L. Ferariu, A. Burlacu, and T. Sauciuc, "Early fusion based CNN architecture for visual servoing systems," MMAR 2022.
- **P. Botezatu**, L. Ferariu, A. Burlacu, and T. Sauciuc, "Visual feedback control using CNN based architecture with input data fusion," ICSTCC 2022.
- **P. Botezatu**, L.-E. Ferariu, and A. Burlacu, "Enhancing visual feedback control through early fusion deep learning," Entropy, vol. 25, no. 10, p. 1378, 2023.
- **P. Botezatu**, A.-I. Iancu, A. Burlacu, "Hybrid Deep Learning Framework for Eye-in-Hand Visual Control Systems", acceptată la Robotics, 2025.

Capitolul 6 introduce o arhitectură hibridă de control vizual ce combină SuperPoint cu algoritmi de control bazat pe cuaternioni duali. Sunt prezentate experimente în timp real care validează eficiența metodei în medii dinamice și complexe.

Capitolul 7 oferă concluzii generale și discută impactul abordărilor propuse, subliniind direcții promițătoare pentru cercetări viitoare în sisteme inteligente bazate pe învățare profundă.

Capitolul 2

Fundamente teoretice și aplicații ale rețelelor neuronale convoluționale

Acest capitol prezintă o introducere în domeniul Învățării profunde, cu accent pe Rețelele neuronale Convoluționale (CNN) și componentele lor esențiale. Se analizează două arhitecturi CNN importante utilizate în această teză, AlexNet și ResNet, evidențiind structura și principiile lor de funcționare. În final, se examinează aplicațiile recente ale CNN-urilor în Servoing Vizual și provocările actuale din domeniu.

2.1 Fundamente ale Rețelelor Neuronale Convoluționale și arhitecturi principale

Rețelele neuronale artificiale (ANN) sunt sisteme de calcul distribuite inspirate din rețelele neuronale biologice, compuse din unități interconectate (neuroni) legate prin conexiuni ponderate. Aceste ponderi sunt ajustate în timpul procesului de învățare supervizată pentru a minimiza eroarea dintre ieșirea prezisă și cea reală.

Rețelele neuronale Convoluționale (CNN) reprezintă un tip specializat de ANN conceput pentru procesarea imaginilor, exploatând structura lor spațială. În arhitectura generală a unui CNN (Figura 2.1), imaginea de intrare este reprezentată ca un tensor tridimensional $W \times H \times D$, unde W este lățimea, H înălțimea și D adâncimea (numărul de canale, 3 pentru imagini RGB).

Extracția caracteristicilor se realizează prin straturi convoluționale, funcții de activare și straturi de pooling. În straturile convoluționale, filtre învățabile (kernels) procesează datele de intrare. După straturile convoluționale se aplică o funcție de activare, cel mai frecvent

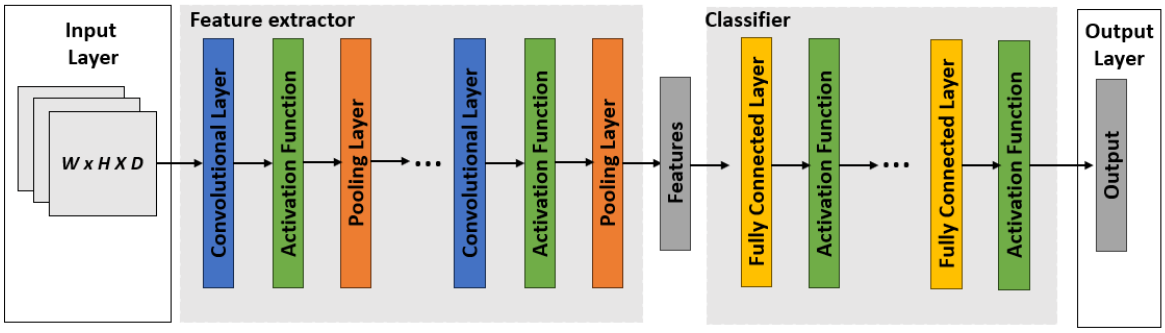

Fig. 2.1 Arhitectură generală a unui CNN

Unitatea Liniară Rectificată (ReLU) [9]. Straturile de pooling reduc dimensiunea spațială a hărților de caracteristici. Cele mai comune operații sunt max-pooling și average pooling.

Clasificatorul utilizează straturi complet conectate pentru a integra caracteristicile și a face predicții, furnizând fie o distribuție de probabilitate pentru clasificare, fie valori numerice pentru regresie.

Două arhitecturi esențiale pentru această teză sunt AlexNet [10] și ResNet-50 [11].

AlexNet (Figura 2.2 - stânga) este o rețea importantă în domeniu, datorită multitudinii de aplicații în care se folosește. Arhitectura sa constă din cinci straturi convoluționale pentru extracția caracteristicilor și trei straturi complet conectate pentru clasificare. O contribuție majoră a AlexNet este utilizarea funcției ReLU, care a atenuat problema gradientului care dispare și a permis antrenarea eficientă a arhitecturilor mai profunde [12].

ResNet-50 (Figura 2.2 - dreapta) este a doua arhitectură folosită în această teză. Rețeaua a introdus un design modular și scalabil care permite antrenarea rețelelor mult mai profunde. Inovația sa principală constă în conexiunile reziduale din fiecare bloc, care permit rețelei să învețe mapări reziduale în locul transformărilor directe.

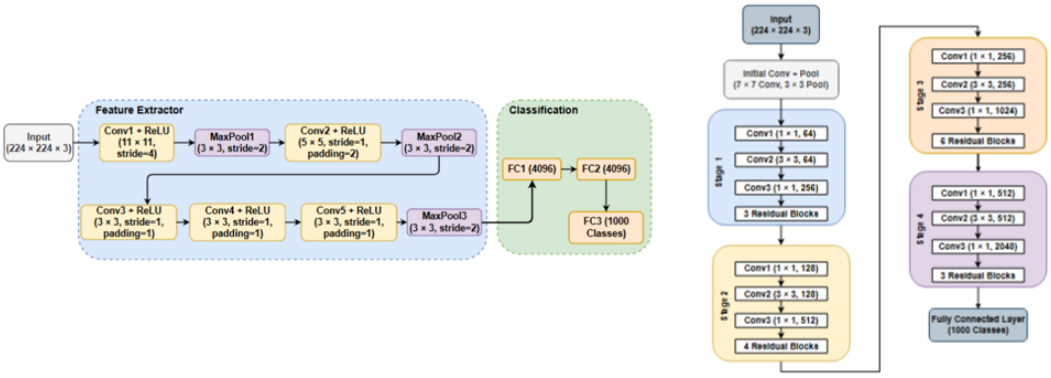

Fig. 2.2 Arhitectura AlexNet (stânga) și ResNet-50 (dreapta)

Capitolul 3

Probleme Deschise ale Aplicațiilor Complexe

3.1 Probleme deschise în clasificarea suprafeței drumurilor

Clasificarea suprafețelor rutiere (RSC) reprezintă o componentă esențială în sistemele autonome de analiză a drumurilor, permițând vehiculelor să înțeleagă caracteristicile suprafeței de rulare. Aceasta implică identificarea și categorizarea suprafețelor rutiere în funcție de compoziția materialului (precum asfalt, beton, pavaj sau pământ) și condițiile de mediu (precum uscat, umed, înghețat sau acoperit de zăpadă). Pentru vehiculele autonome, o clasificare precisă oferă informații contextuale esențiale care afectează direct dinamica vehiculului, planificarea traseului și parametrii de siguranță. În ciuda importanței sale și a progreselor considerabile realizate în ultimii ani prin abordări de învățare profundă, RSC continuă să se confrunte cu provocări semnificative care limitează implementarea sa practică în scenarii reale. Sistemele RSC actuale se bazează predominant pe informații vizuale din camere, având limitări inerente în medii dificile precum iluminare slabă sau condiții meteorologice adverse. Variabilitatea geografică a condițiilor rutiere prezintă provocări substanțiale pentru dezvoltarea sistemelor RSC aplicabile la nivel global. Metodele curente abordează clasificarea suprafețelor ca o problemă de analiză independentă cadru cu cadru, ignorând natura continuă și tranzițională a condițiilor suprafeței rutiere. Majoritatea abordărilor RSC se concentrează pe categorizarea largă a suprafețelor în câteva clase limitate, iar performanța sistemelor RSC în diferite condiții de mediu și variații sezoniere rămâne insuficient explorată. Cercetările viitoare în RSC ar trebui să se concentreze pe dezvoltarea soluțiilor integrate care abordează aceste provocări multiple simultan. Abordările de fuziune la intrare a datelor, în special pentru datele de imagine multi-canal, oferă o direcție promițătoare. Prin combinarea

diferitelor modalități vizuale (RGB, infraroșu, adâncime, termic) la nivelul de intrare, fuziunea la intrare a datelor permite modelelor să învețe reprezentări comune care sunt în mod inerent mai robuste la variațiile de mediu, schimbările temporale și diferențele geografice. Această abordare ar putea rezolva provocarea percepției multimodale prin integrarea fluxurilor complementare de informații vizuale și ar facilita clasificarea fină prin captarea indiciilor vizuale subtile între modalități, necesitând totodată mai puțină putere de calcul comparativ cu metodele de fuziune la iesire a datelor.

3.2 Probleme deschise în servoing-ul vizual

Sistemele de control cu feedback vizual, cunoscute în mod obișnuit sub numele de servoing vizual (VS), integrează vederea artificială cu teoria controlului pentru a permite sarcini precise de manipulare și navigare în robotică. Folosind informații vizuale în timp real, aceste sisteme își pot ajusta dinamic acțiunile ca răspuns la schimbările de mediu, îmbunătățindu-și robustețea și versatilitatea.

Scopul unui sistem VS este de a minimiza o eroare $e(t)$ definită în [6] ca:

$$e(t) = s(m(t), a) - s^* \tag{3.1}$$

unde $m(t)$ este un set de măsurători de imagine, $s(m(t), a)$ este un vector de k caracteristici vizuale, în care a este un set de parametri care reprezintă o potențială cunoaștere suplimentară despre sistem (de exemplu, parametri intrinseci ai camerei), iar vectorul s^* reprezintă valorile dorite ale caracteristicilor (de exemplu, poziția în planul imaginii). Există trei abordări principale pentru definirea lui s, reprezentate de cele trei abordări clasice ale controlului de servoing vizual:

- Servoing Vizual Bazat pe Imagine (IBVS) definește s pentru a reprezenta un set de caracteristici disponibile imediat în imagine (de exemplu, coordonatele planului imaginii ale unor puncte urmărite)
- Servoing Vizual Bazat pe Poziție (PBVS) definește s prin intermediul unui set de parametri 3D, care trebuie estimați din imagine
- Servoing Vizual Hibrid (Hybrid VS) care combină elemente din IBVS și PBVS pentru a valorifica punctele forte ale fiecărei abordări, atenuând în același timp slăbiciunile respective

IBVS este o abordare directă în servoing vizual în care acțiunile de control sunt derivate direct din caracteristicile vizuale prezente în planul imaginii. În IBVS, caracteristicile vizuale

s sunt definite în mod tipic ca coordonatele în pixeli ale unor puncte specifice sau modele urmărite în imaginile succesive capturate de cameră.

Servoing Vizual Bazat pe Poziție (PBVS) este o metodă de control cu feedback vizual în care sistemul se bazează pe poziția 3D estimată a obiectului țintă în raport cu camera pentru a calcula comenzile de control. Spre deosebire de IBVS, PBVS operează în spațiul cartezian, oferind control explicit asupra poziției și orientării robotului.

Servoing Vizual Hibrid (HVS) a fost dezvoltat pentru a depăși limitările abordărilor tradiționale, combinând avantajele IBVS [13] și PBVS [14]. Prin integrarea feedback-ului 2D și 3D într-un cadru unificat, HVS elimină dezavantajele individuale ale acestor metode [15].

Progresele recente au integrat tehnici de învățare profundă pentru a gestiona medii dinamice. Studiul din [5], utilizând peste 800.000 de încercări de prindere, a demonstrat că CNN-urile pot realiza coordonare mână-ochi end-to-end fără calibrare manuală. Similar, [16] a eliminat necesitatea extracției de caracteristici prin antrenarea unui CNN end-to-end pentru IBVS. Integrarea CNN-urilor cu Servoing Vizual Direct (DVS) permite estimarea directă a matricelor de interacțiune din datele brute ale imaginii, îmbunătățind robustețea și abordând probleme precum sensibilitatea la calibrare și minimele locale [17]. Această combinație reprezintă o schimbare de paradigmă, oferind soluții eficiente și robuste atât în scenarii structurate, cât și nestructurate [18, 19].

Capitolul 4

Arhitecturi CNN cu Fuziunea Datelor la Intrare

Acest capitol explorează arhitecturile de rețele neuronale convoluționale (CNN) cu fuziune la intrare a datelor și aplicațiile lor în clasificarea suprafețelor rutiere. Se analizează conceptele teoretice ale fuziunii datelor și se demonstrează, prin studii de caz detaliate, cum aceasta poate îmbunătăți acuratețea clasificării în aplicații din lumea reală.

4.1 Fuziunea Datelor în Rețelele neuronale

Fuziunea datelor integrează informații din surse sau modalități diverse pentru a depăși limitările intrărilor unimodale și pentru a exploata caracteristici complementare. În funcție de etapa în care sunt combinate datele, strategiile de fuziune pot fi clasificate în trei categorii principale: Fuziunea târzie (nivel de decizie) - combină rezultatele după ce fiecare modalitate a fost procesată independent; Fuziunea intermediară (hibridă) - combină date în straturi intermediare ale rețelei; Fuziunea datelor la intrare (nivel de intrare) - integrează fluxurile de date în cea mai timpurie etapă, prin concatenarea canalelor brute.

Această lucrare se concentrează pe fuziunea datelor la intrare, care prezintă avantaje precum eficiența computațională și capacitatea de a învăța interacțiuni între modalități de la început, deși poate fi mai susceptibilă la dezechilibre și incompatibilități între tipurile de date.

4.2 Arhitectura CNN cu Fuziune la Intrare a Datelor

Arhitectura propusă, ilustrată în Figura 4.1, integrează hărți suplimentare alături de imaginile RGB la nivelul intrării, permițând rețelei să învețe automat relațiile întrea imaginea RGB și harta suplimentară. Acest concept este aplicat atât pentru clasificarea suprafețelor rutiere, cât și pentru controlul vizual cu feedback în robotică.

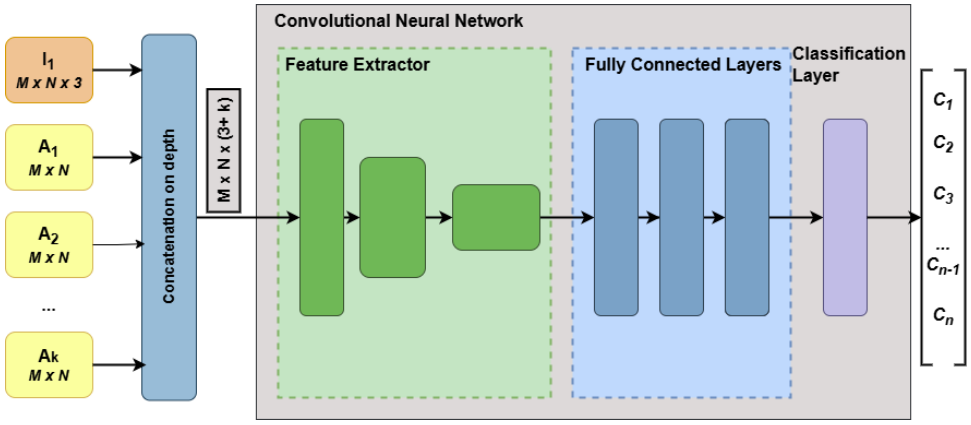

Fig. 4.1 Arhitectura generală a unui CNN bazat pe fuziunea la intrare a datelor

Intrarea rețelei este formată prin concatenarea imaginii RGB (I_1) cu k hărți suplimentare (A_1 până la A_k), rezultând un tensor de intrare cu dimensiune crescută. Această abordare permite o implementare eficientă în timp real, deoarece modificările se concentrează pe augmentarea intrării, nu pe creșterea complexității rețelei.

Pentru adaptarea arhitecturilor pre-antrenate la fuziunea propusă, primul strat convoluțional este modificat pentru a accepta canale de intrare suplimentare, în timp ce restul rețelei rămâne neschimbat pentru a beneficia de capacitățile sale de extracție a caracteristicilor.

4.3 Studiu de Caz: Clasificarea Suprafețelor Rutiere

Clasificarea suprafețelor rutiere este esențială pentru conducerea autonomă și sistemele de transport inteligente, influențând direct siguranța, navigația și luarea deciziilor. Abordările tradiționale care se bazează doar pe imagini RGB pot fi limitate în condiții de iluminare variabilă sau vreme adversă, iar fuziunea timpurie oferă o soluție prin integrarea de informații complementare.

4.3.1 Caracteristici Vizuale pentru Clasificarea Suprafețelor

Pentru îmbunătățirea clasificării suprafețelor rutiere, au fost analizate două tipuri principale de caracteristici vizuale. Primele au fost hărți în tonuri de gri care reduc complexitatea imaginilor RGB la o reprezentare de intensitate cu un singur canal, evidențiind diferențele de textură între suprafețe, cum se observă în Figura 4.2 - b. Al doilea tip de caracteristici vizuale a fost reprezentat de hărțile de gradient care accentuează schimbările de intensitate spațială, evidențiind marginile și tranzițiile din suprafețele rutiere, cum se vede în Figura 4.2 - c

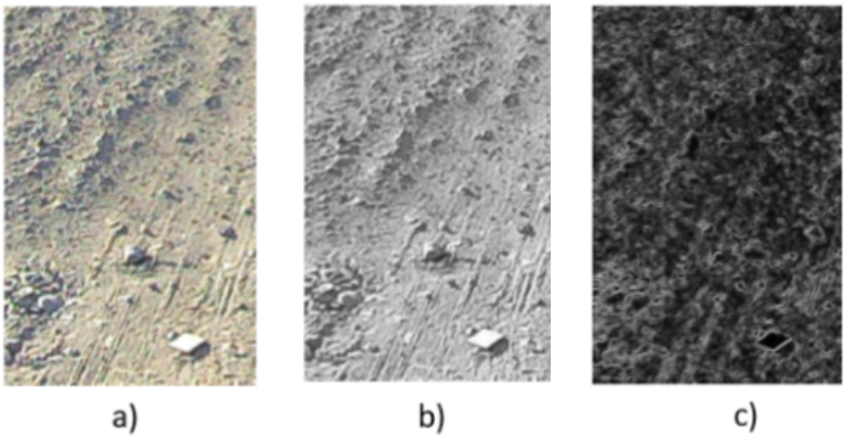

Fig. 4.2 Exemplu de imagine RGB și hărțile corespunzătoare în tonuri de gri: a) Imagine RGB; b) Imagine în tonuri de gri; c) Imagine de tip gradient

4.3.2 Evaluare Experimentală

Evaluarea a fost efectuată în două scenarii: unul balansat din punctul de vedere al claselor și unul nebalansat, pentru a testa robustețea abordării propuse în condiții variate.

Scenariu cu setul de date balansat

Primul scenariu utilizează un subset balansat cu opt clase de suprafețe rutiere din setul de date RSCD, prezentate în Figura 4.3. Pentru fiecare clasă s-a utilizat un număr egal de imagini pentru a asigura un proces de învățare corect.

Rezultatele comparative pentru cele trei configurații testate (fără fuziune timpurie, fuziune timpurie cu tonuri de gri și fuziune timpurie cu gradient) sunt prezentate în Figura 4.4. Fuziunea timpurie cu tonuri de gri a atins cea mai mare acuratețe (0,9293), urmată de fuziunea timpurie cu gradient (0,9247) și configurația fără fuziune timpurie (0,9076).

Fig. 4.3 Scenariu balansat: cele opt clase selectate

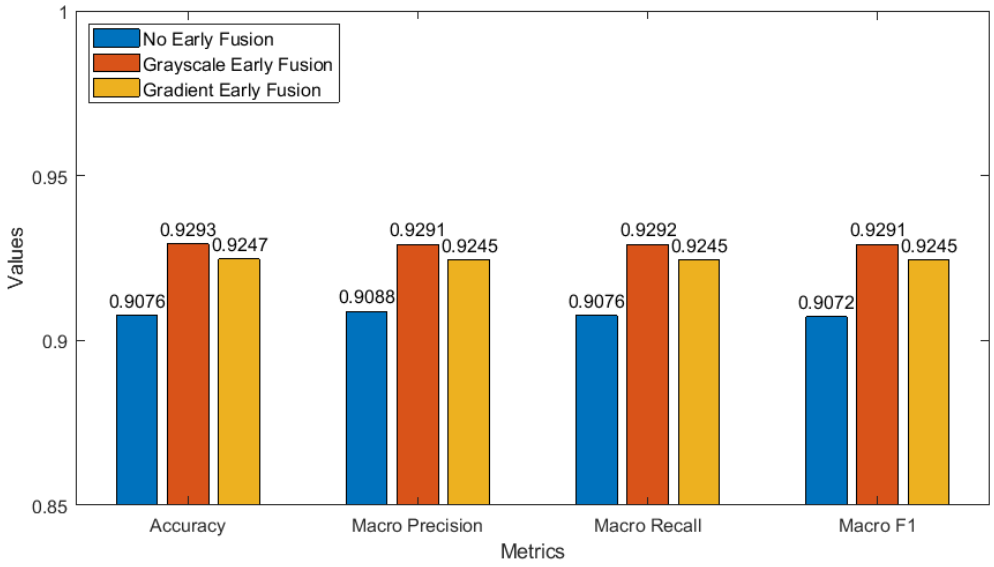

Fig. 4.4 Metrici generale pentru cele trei configurații în scenariul balansat

Scenariu Nebalansat

Al doilea scenariu utilizează un subset nebalansat cu zece clase din setul de date RSCD, prezentat în Figura 4.5, reflectând problema cozii lungi din seturile de date din lumea reală.

Rezultatele comparative pentru acest scenariu sunt prezentate în Figura 4.6. Configurația cu fuziune timpurie cu tonuri de gri a obținut din nou cele mai bune rezultate, cu o acuratețe de 0,8688, demonstrând robustețea abordării chiar și în condiții dezechilibrate.

Fig. 4.5 Scenariu nebalansat: cele zece clase selectate

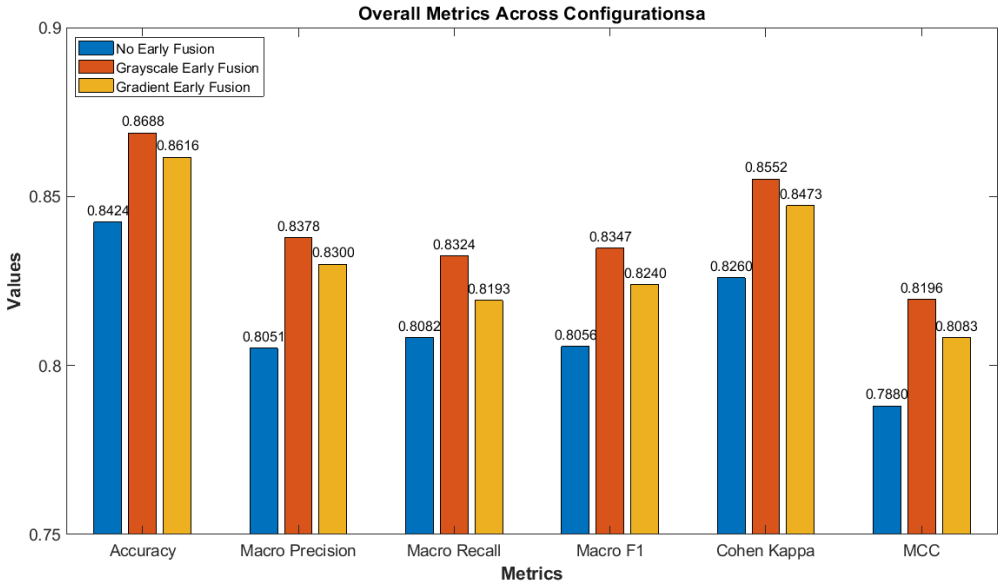

Fig. 4.6 Metrici generale pentru cele trei configurații în scenariul nebalansat

Rezultatele experimentale demonstrează că abordările bazate pe fuziune la intrare a datelor îmbunătățesc semnificativ performanța clasificării suprafețelor rutiere, atât în scenarii echilibrate, cât și dezechilibrate. Fuziunea la intrare cu tonuri de gri oferă cele mai bune rezultate, datorită capacității sale de a capta variații subtile de textură care sunt esențiale pentru distingerea diferitelor tipuri de suprafețe rutiere.

În scenariul nebalansat, abordarea cu fuziune a datelor demonstrează o robustețe superioară față de metoda tradițională fără fuziune, evidențiind potențialul său pentru aplicații

din lumea reală. Beneficiile fuziunii timpurii vin cu un cost computațional minim, deoarece modificările sunt concentrate la nivelul intrării, nu în arhitectura rețelei

Acest capitol stabilește fundamentul teoretic și practic pentru aplicarea fuziunii timpurie în controlul vizual cu feedback, care va fi explorat în capitolele următoare. Integrarea caracteristicilor vizuale complementare îmbunătățește capacitatea rețelelor neuronale de a extrage informații relevante, conducând la sisteme mai robuste și mai precise în aplicații de conducere autonomă și robotică.

Capitolul 5

Control vizual cu feedback bazat pe Învățare Profundă

Acest capitol oferă o explorare cuprinzătoare a controlului vizual cu feedback bazat pe învățare profundă prin integrarea indicilor tradiționali de servocontrol vizual cu arhitecturi CNN avansate cu fuziune timpurie. Pornind de la conceptele stabilite în servocontrolul vizual, sunt introduse multiple caracteristici vizuale complementare, inclusiv hărți de segmentare bazate pe regiuni, hărți cu puncte de interes și hărți cu momente de imagine, demonstrând cum acestea pot fi combinate la nivelul de intrare pentru a îmbunătăți precizia controlului pentru un sistem robotic cu 6 grade de libertate.

1. **Hărți de segmentare a imaginilor** - Fiecare regiune corespunzătoare fundalului sau obiectului este etichetată cu intensitatea sa medie, oferind avantaje precum rezistența la variațiile de iluminare și facilitarea potrivirii obiectelor.

2. **Hărți cu puncte de interes** - Utilizează operatori SURF și BRISK pentru a identifica locații cheie robuste la schimbări de scală, rotație și calitate a imaginii.

3. **Hărți cu momente de imagine** - Folosește descriptori matematici care captează caracteristici cheie precum poziția, orientarea și forma obiectelor din imagine.

5.1 Rezultate experimentale offline

Evaluarea offline a fost realizată utilizând setul de date VS [4], colectat cu un braț robotic Kinova Gen3, conținând scene cu un singur obiect (Scenariul Experimental 1 - ES1) și cu mai multe obiecte (Scenariul Experimental 2 - ES2). Pentru fiecare scenariu au fost generate 30.000 de combinații de configurații curente și finale, împărțite în seturi de antrenare (70%),

validare (15%) și testare (15%). În cele ce urmează se prezintă rezultatele pentru al doilea scenariu de testare.

Pentru ES2, s-au dezvoltat mai multe arhitecturi CNN:M_1^{ES1} - fără date de fuziune timpurie, M_2^{ES1} - cu hărți de segmentare bazate pe regiuni, M_3^{ES1} - cu puncte de interes BRISK, M_4^{ES1} - cu puncte de interes SURF.

Rezultatele, prezentate în Figura 5.1 au arătat că hărțile de segmentare au un impact mai mare asupra arhitecturii neurale decât hărțile cu puncte de interes în scenele complexe. Punctele de interes par să fie mai eficiente în scene experimentale mai simple, deoarece captează caracteristici distinctive ale unui singur obiect. În cazul scenelor complexe cu mai multe obiecte, hărțile de segmentare s-au dovedit mai avantajoase, deoarece izolează obiectele individuale de fundal.

Convergenta vitezelor, prezentată în Figura 5.1, compară vitezele generate de modele cu cele așteptate și cele calculate folosind legea de control PBVS. Se poate observa că modelele neurale prezic viteze mai apropiate de valorile de referință comparativ cu cele generate de legea de control din PBVS, evidențiind avantajele arhitecturii de învățare profundă cu fuziune timpurie pentru predicția și controlul mișcării.

Pentru arhitecturile bazate pe momente de imagine unghiulare, s-au considerat diverse configurații: Folosind ecuațiile lui Chaumette vs. cele ale lui Tahri, cu sau fără fuziunea cu imagini RGB iar in final, cu diferite dimensiuni și suprapuneri ale celulelor.

Compararea rezultatelor a arătat că punctele de interes și hărțile de segmentare au obținut aproximări mai precise comparativ cu hărțile de momente de imagine. Abordarea de antrenare în două etape a produs valori MSE îmbunătățite față de antrenarea exclusiv cu hărți de momente de imagine unghiulare.

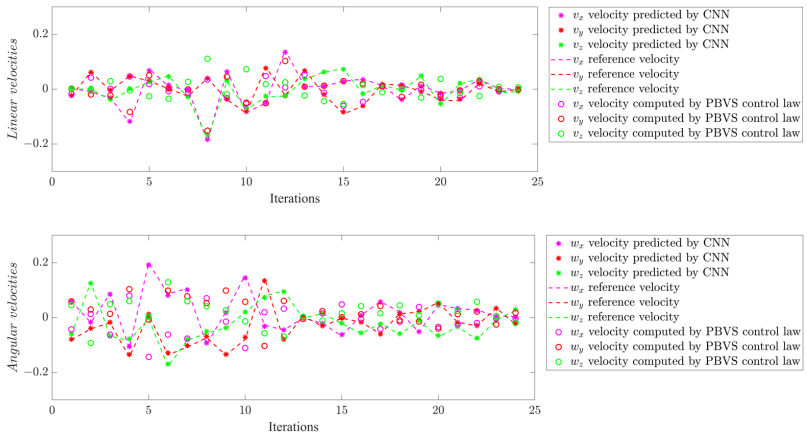

Fig. 5.1 Viteze liniare (sus) și unghiulare (jos) generate pentru Scena Experimentală Complexă, ES2

5.2 Rezultate experimentale online

Pentru evaluarea online, cele mai bune modele derivate din evaluarea offline au fost integrate în scenariul de testare în timp real și ajustate folosind un set de date achiziționat în timp real.

Pregătirea setului de date pentru teste online

Datele au fost colectate folosind un robot colaborativ UR5 echipat cu o cameră stereo ZED Mini montată pe efector. Setul de date a fost creat utilizând două abordări. Prima a implicat o achiziție manuală unde elementul terminal este deplasat manual în configurații aleatorii, unde imaginile și pozele sunt achiziționate instantaneu iar a doua o generare sintetică, unde imaginile au fost generate sintetic folosind perturbări aleatorii ale pozei de referință și transformări homografice.

Figura 5.2 arată imaginea de referință utilizată pentru generarea sintetică, iar Figura 5.3 prezintă două imagini aleatorii generate din această imagine de referință.

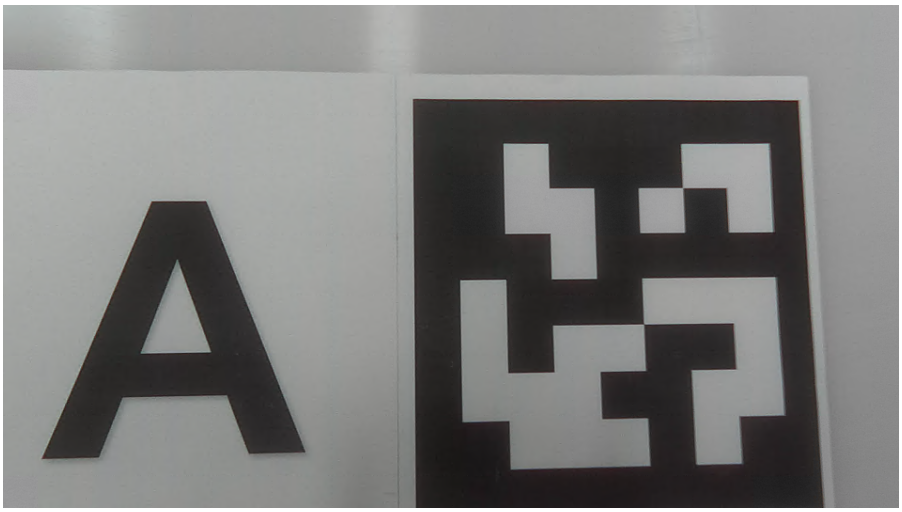

Fig. 5.2 Exemplu de imagine de referință utilizată pentru generarea sintetică

Analiza scenariului de control în timp real

Pentru evaluarea online a fost considerat scenariul prezentat în Figura 5.4.

Analiza convergenței vitezei, ilustrată în Figura 5.5 pentru primul scenariu de testare, arată că modelul cu fuziune timpurie (stânga) produce rezultate mai netede cu oscilații reduse și atinge o stare de viteză aproape de zero mai rapid decât modelul fără fuziune timpurie (dreapta).

Fig. 5.3 Exemplu de două imagini sintetice generate aleatoriu din imaginea de referință

Fig. 5.4 Scenarii de testare online considerate pentru sarcina VS

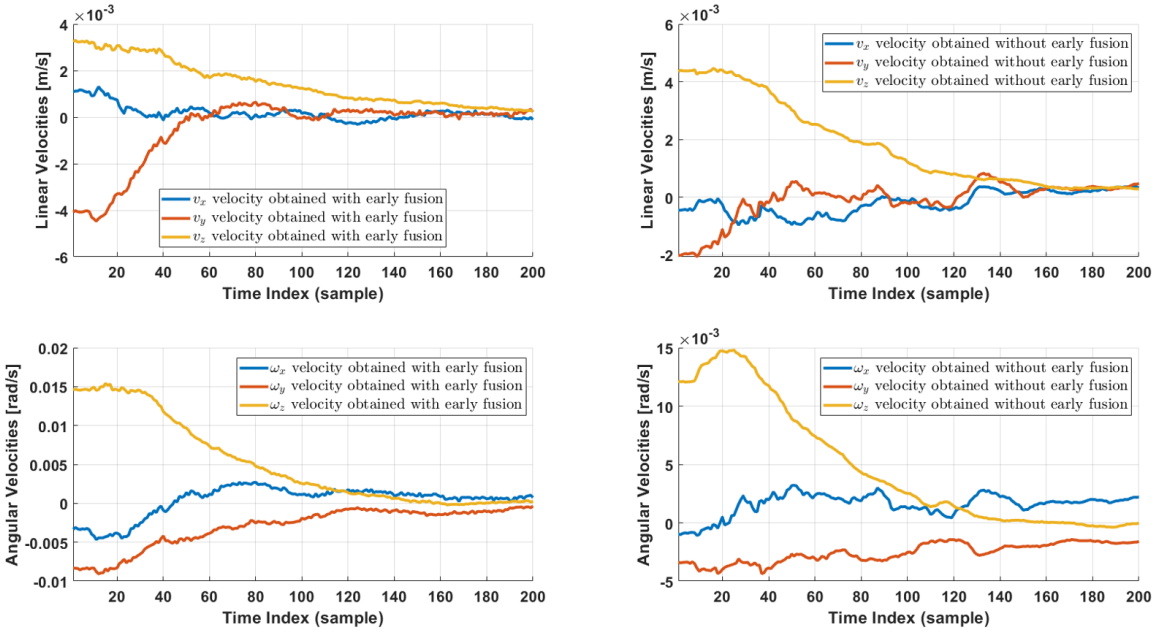

Fig. 5.5 Analiza vitezei în primul scenariu de testare online. Stânga: Date obținute cu fuziune timpurie. Dreapta: Date obținute fără fuziune timpurie

Analiza metricilor la nivel de pixel a fost realizată pentru a evalua alinierea la nivel de imagine între configurațiile curente și dorite. Figura 5.6 - stanga ilustrează evoluția erorii

medii pătratice (MSE), erorii mediane pătratice și deviației standard pentru fiecare imagine din primul scenariu de testare. Modelul ResNet cu fuziune timpurie (arătat în roșu) atinge în mod consistent valori mai mici în toate aceste metrici comparativ cu modelul de bază (arătat în albastru).

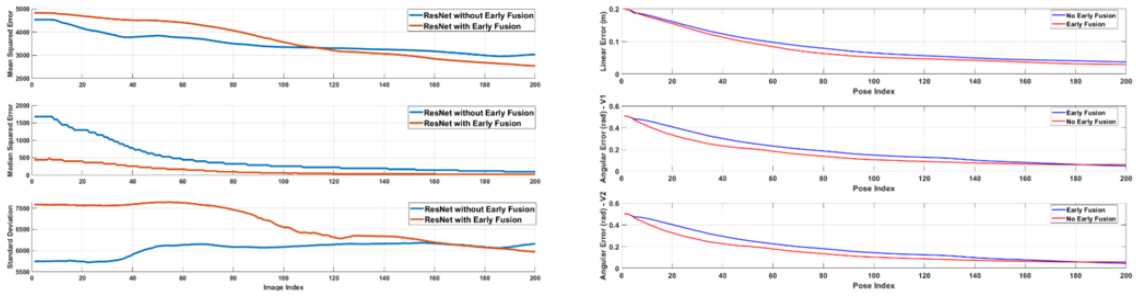

Fig. 5.6 Comparația metricilor de eroare la nivel de pixel pentru primul scenariu de testare online

Analiza erorii de poziție, prezentată în Figura 5.6 - dreapta, pentru primul scenariu de testare, demonstrează că arhitectura ResNet cu fuziune timpurie atinge o convergență mai rapidă atât pentru poziție, cât și pentru orientare, menținând niveluri de eroare mai scăzute pe parcursul evaluării. În contrast, modelul de bază converge mai lent și prezintă oscilații notabile, în special în a doua metrică de eroare unghiulară.

Rezultatele acestor analize sugerează că metoda de fuziune timpurie, realizată prin încorporarea directă a hărților de puncte caracteristice SURF în straturile inițiale ale CNN, îmbunătățește semnificativ capacitatea modelului de a generaliza robust. Informațiile suplimentare furnizate de hărțile SURF permit rețelei să gestioneze mai bine configurațiile noi, subliniind importanța integrării în stadiile timpurii pentru o mai bună generalizare a modelului.

Arhitectura cu fuziune timpurie se dovedește deosebit de avantajoasă în scenarii de testare online, unde demonstrează: o convergență mai rapidă și mai stabilă a vitezelor, erori reduse la nivel de pixeli, o aliniere mai precisă a poziției și orientării precum si capacitate îmbunătățită de a gestiona configurații necunoscute

Aceste rezultate confirmă că integrarea caracteristicilor vizuale tradiționale în arhitecturile moderne de învățare profundă poate îmbunătăți semnificativ precizia și robustețea controlului vizual cu feedback, deschizând calea către aplicații avansate de servocontrol vizual în robotică.

Capitolul 6

Control vizual cu feedback bazat pe date

Acest capitol explorează o abordare bazată pe date pentru controlul vizual cu feedback, utilizând rețeaua neuronală SuperPoint [20] pentru detecția și descrierea punctelor de interes și integrarea acestora într-un algoritm de control bazat pe algebra Lie [21]. Această metodă permite calcularea directă a vitezelor robotului din coordonatele carteziene 3D ale punctelor de interes potrivite.

6.1 Arhitectura SuperPoint

SuperPoint reprezintă un avans semnificativ în domeniul detecției și descrierii punctelor de interes, folosind învățarea profundă într-o rețea neuronală complet convoluțională. Spre deosebire de metodele clasice, SuperPoint utilizează o strategie de antrenare auto-supravegheată numită Adaptare Homografică, eliminând necesitatea etichetării manuale extensive.

Arhitectura SuperPoint constă din trei componente interconectat: un codificator comun bazat pe arhitectura convoluțională VGG, un decodor pentru puncte de interes care generează o hartă de probabilitate si un decodor pentru descriptori care calculează descriptori robuști pentru potrivirea între imagini.

Figurile 6.1 și 6.2 prezintă exemple de potriviri realizate de rețea pe o scenă simplă (cu markeri) și una complexă (cu obiecte multiple), evidențiind cele mai importante șase potriviri pentru punctele detectate, valoare aleasă euristic pentru exemplificare.

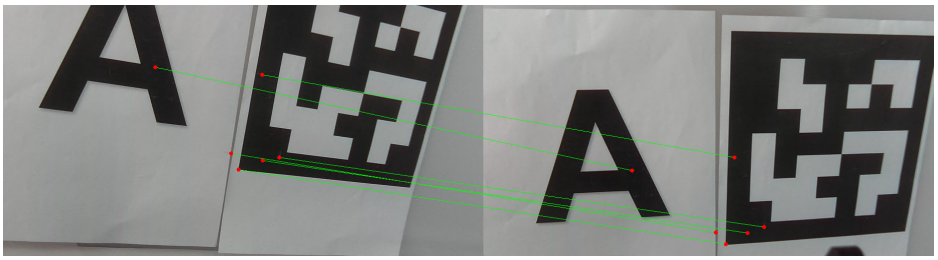

Fig. 6.1 Exemplu de potriviri SuperPoint pe o scenă simplă

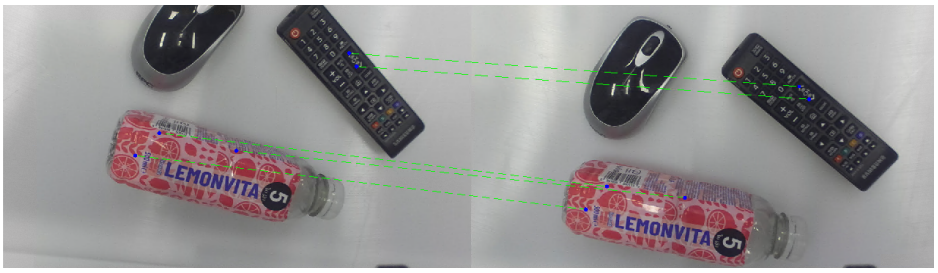

Fig. 6.2 Exemplu de potriviri SuperPoint pe o scenă complexă

Arhitectura sistemului

Metoda propusă utilizează potrivirile de puncte cheie generate de SuperPoint și le integrează într-un algoritm de control bazat pe algebra Lie a cuaternionilor duali. Arhitectura este structurată în trei straturi interconectate, ilustrată în Figura 6.3:

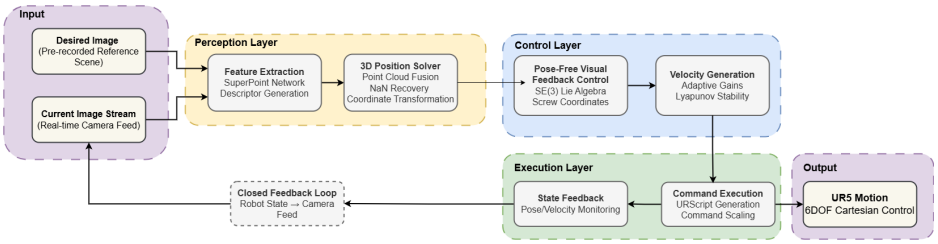

Fig. 6.3 Arhitectura de servocontrol vizual bazată pe SuperPoint

- **Stratul de Percepție**: SuperPoint detectează și potrivește puncte de interes robuste între imaginea de referință și fluxul camerei în timp real. Aceste puncte sunt asociate cu pozițiile 3D corespunzătoare extrase din norul de puncte, furnizând coordonate carteziene 3D fiabile pentru algoritmul de control.
- **Stratul de Control**: Pozițiile 3D derivate din stratul de percepție sunt utilizate pentru calcularea vitezelor liniare și unghiulare ale efectorului final. O abordare fără poziție minimizează eroarea 3D punct-la-punct și asigură convergența către configurația dorită.

- **Stratul de Execuție**: Robotul execută comenzile de viteză calculate, transformând comenzile teoretice în mișcări robotice reale prin URScript, cu monitorizare și feedback în timp real.

6.2 Rezultate experimentale

Evaluarea experimentală a arhitecturii propuse a fost realizată pe două scene: scena simplă SuperPoint (SSS) cu markeri și scena complexă SuperPoint (SCS) cu trei obiecte, ilustrate în Figurile 6.4 și 6.5.

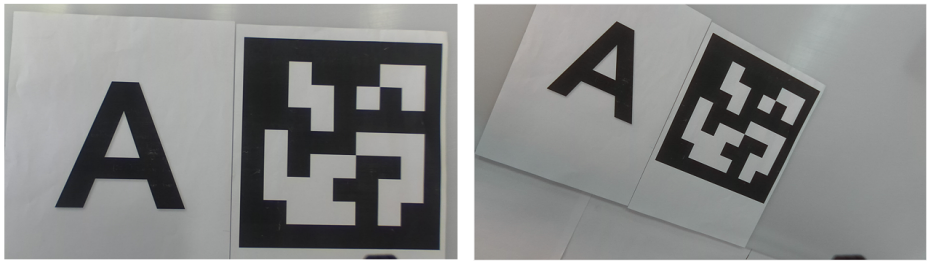

Fig. 6.4 Configurațiile inițială (dreapta) și dorită (stânga) pentru Scena Simplă SuperPoint

Fig. 6.5 Configurațiile inițială (dreapta) și dorită (stânga) pentru Scena Complexă SuperPoint

Pentru fiecare scenă, au fost testate trei configurații diferite de parametri de control: o configurație echilibrată: $\alpha_P = \alpha_G = 0.1$, o configurație cu : $\alpha_P = 0.1, \alpha_G = 0.05$ si o configurație cu : $\alpha_P = 0.05, \alpha_G = 0.1$.

Rezultate pentru Scena Simplă

Analiza convergenței vitezei

Experimentele pentru scena simplă demonstrează performanța sistemului în stabilizarea vitezelor liniare și unghiulare. Figura 6.6 ilustrează rezultatele pentru configurația echilibrată,

unde componentele de viteză liniară prezintă inițial un comportament oscilatoriu care se atenuează progresiv pe măsură ce efectorul final se apropie de configurația dorită.

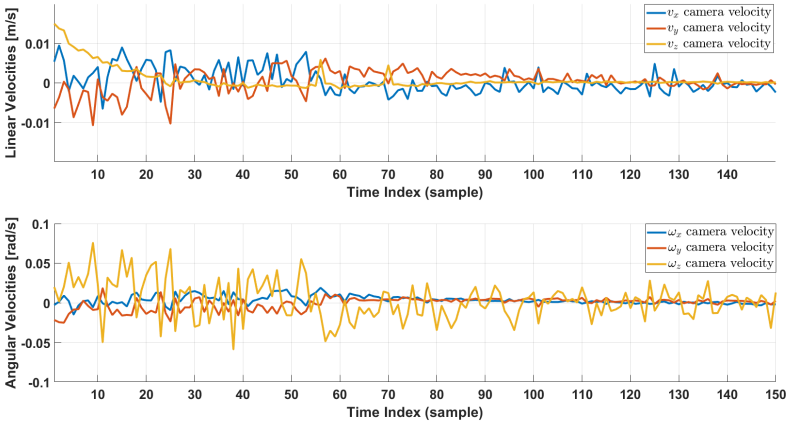

Fig. 6.6 Convergența vitezei liniare și unghiulare pentru Scena Simplă SuperPoint cu ponderi $\alpha_P = 0.1$ și $\alpha_G = 0.1$

Analiza comparativă a celor trei configurații relevă că in primul rand ca configurația echilibrată oferă o performanță generală bună cu oscilații moderate. De asemenea demonstraza ca reducerea câștigului centroidului duce la vârfuri de viteză mai pronunțate și stabilizare întârziată dar si ca reducerea câștigului de deplasare a punctului produce oscilații inițiale ale vitezei unghiulare mai mari care se atenuează mai rapid.

Analiza metricilor la nivel de pixel

Metricile de eroare la nivel de pixel oferă informații complementare despre eficiența convergenței vizuale. Figura 6.7 - stanga prezintă rezultatele pentru configurația echilibrată, demonstrând o îmbunătățire semnificativă în alinierea imaginii.

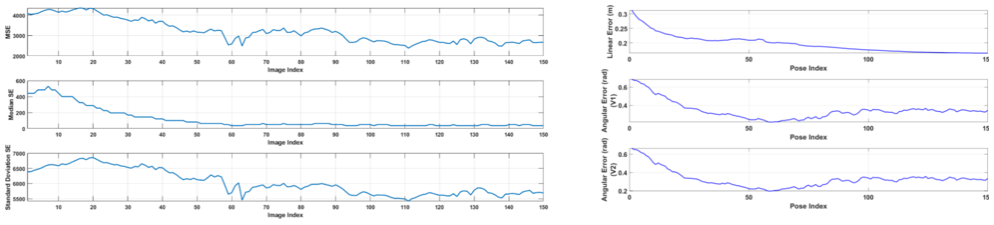

Fig. 6.7 Metrici de eroare la nivel de pixel pentru Scena Simplă SuperPoint cu ponderi $\alpha_P = 0.1$ și $\alpha_G = 0.1$

Eroarea Mediană Pătratică demonstrează o convergență rapidă, indicând o aliniere excelentă în majoritatea imaginii. Configurația cu câștig de deplasare a punctului redus realizează cea mai bună convergență a erorii mediane, în ciuda valorilor mai mari ale mediei și deviației standard. Compromisurile fundamentale în proiectarea controlerului: ponderile echilibrate oferă o convergență generală consistentă, în timp ce configurațiile asimetrice pot prioritiza fie alinierea globală, fie precizia specifică

Analiza erorii de poziție

Analiza erorii de poziție, ilustrată în Figura 6.7 - dreapta pentru configurația echilibrată, arată o convergență eficientă atât în aspectele poziționale, cât și în cele orientaționale.

Comparația între cele trei configurații arată că toate configurațiile realizează o reducere similară a erorii liniare (aproximativ 45%), configurația cu câștig global redus atinge minime unghiulare mai profunde dar mai puțin stabila dar si ca configurația cu câștig de deplasare a punctului redus demonstrează cea mai bună reducere a erorii unghiulare, sugerând că prioritizarea mișcării globale față de corecția individuală a caracteristicilor beneficiază alinierii orientării.

Rezultate pentru Scena Complexă

Rezultatele pentru scena complexă cu multiple obiecte indică câteva diferențe importante față de scena simplă, surprinse în Figura 6.8.

Se observa ca configurația echilibrată demonstrează cea mai eficientă performanță generală pentru scena complexă, cu convergență stabilă după aproximativ 50 de eșantioane. Configurația cu câștig centroid redus oferă o precizie superioară a poziției în ciuda convergenței inițiale mai lente. Configurația cu câștig de deplasare a punctului redus prezintă oscilații persistente, sugerând importanța menținerii unei capacități adecvate de corecție punctuală pentru scene complexe

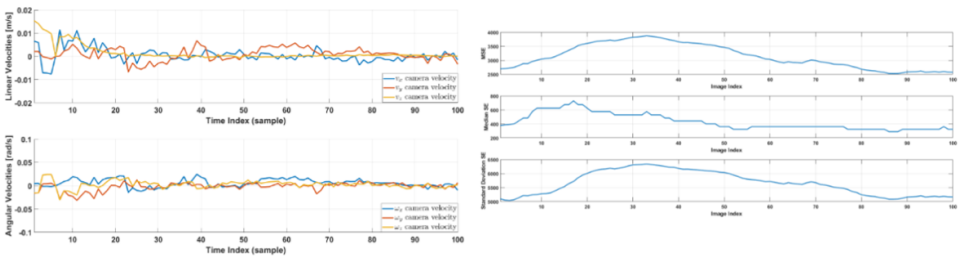

Fig. 6.8 Convergența vitezei și erori la nivel de pixel pentru Scena Complexă

Analiza cuprinzătoare a scenelor simple și complexe demonstrează că performanța sistemului de servocontrol vizual bazat pe SuperPoint este influențată semnificativ de complexitatea scenei și de selecția parametrilor de control. Pentru scena simplă, ponderile echilibrate oferă performanța optimă, în timp ce pentru scena complexă, un câștig centroid redus oferă o precizie superioară a poziției.

Arhitectura propusă demonstrează o robustețe remarcabilă în stabilirea unei corespondențe fiabile a caracteristicilor în diverse condiții vizuale, iar cadrul de control bazat pe cuaternioni duali traduce eficient aceste corespondențe în comenzi de mișcare adecvate. Compromisurile observate între viteza de convergență, stabilitate și precizia finală evidențiază importanța adaptării parametrilor de control la cerințele specifice ale aplicației,precum si ponderi echilibrate pentru scene simple care necesită traiectorii fluide. Această abordare bazată pe date pentru controlul vizual cu feedback oferă un cadru flexibil și robust pentru aplicații de servocontrol vizual, adaptându-se la diverse configurații ale scenei prin utilizarea inteligentă a caracteristicilor vizuale extrase direct din imagini.

Capitolul 7

Concluzii și direcții viitoare de cercetare

În ultimii ani, progresele în Inteligența Artificială și Învățarea Profundă au transformat fundamental sistemele de siguranță auto și capacitățile de manipulare robotică. Obiectivul principal al acestei teze a fost dezvoltarea și evaluarea arhitecturilor de învățare profundă cu fuziune timpurie care îmbunătățesc atât clasificarea suprafețelor rutiere pentru Sistemele Avansate de Asistență a Șoferului (ADAS), cât și controlul cu feedback vizual pentru manipularea robotică. Cercetarea s-a concentrat pe trei direcții principale: (1) integrarea canalelor suplimentare de imagini (grayscale și gradient) alături de date RGB pentru clasificarea robustă a suprafețelor rutiere, (2) dezvoltarea unui framework comprehensiv pentru servoing vizual care combină diverse caracteristici vizuale (hărți de segmentare, puncte de interes și momente angulare), și (3) crearea unui sistem hibrid care integrează framework-ul SuperPoint cu algoritmi de control bazați pe cuaternioni duali pentru manipulare robotică precisă. Experimentele realizate atât în medii simulate, cât și pe platforme robotice reale, confirmă eficacitatea abordărilor propuse și aplicabilitatea lor practică în sistemele autonome din lumea reală.

7.1 Diseminarea rezultatelor

Rezultatele din această teză însumează 4 lucrări de jurnal și 5 lucrări de conferință, după cum urmează.

Lucrări de jurnal:

- **A.-P. Botezatu** and A. Burlacu, "A short review of deep learning methods in visual servoing systems," Bulletin of the Polytechnic Institute of Iasi, Electrical Engineering, Power Engineering, Electronics Section, vol. 69, no. 3, pp. 113–136, 2023.

- **A.-P. Botezatu**, L.-E. Ferariu, and A. Burlacu, "Enhancing visual feedback control through early fusion deep learning," Entropy, vol. 25, no. 10, p. 1378, 2023. (Q2, IF - 2.1)
- **A.-P. Botezatu**, A. Burlacu, and C. Orhei, "A review of deep learning advancements in road analysis for autonomous driving," Applied Sciences, vol. 14, no. 11, p. 4705, 2024. (Q2, IF - 2.5)
- **A.-P. Botezatu**, A.-I. Iancu, A. Burlacu, "Hybrid Deep Learning Framework for Eye-in-Hand Visual Control Systems", Robotics, vol.14, no.5, 2025 (Q2, IF - 2.9)

Lucrări de conferință:

- **A.-P. Botezatu**, L. Ferariu, A. Burlacu, and T. Sauciuc, "Early fusion based CNN architecture for visual servoing systems," in 2022 26th International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 199–204, IEEE, 2022. (WoS)
- **A.-P. Botezatu**, L. Ferariu, A. Burlacu, and T. Sauciuc, "Visual feedback control using CNN based architecture with input data fusion," in 2022 26th International Conference on System Theory, Control and Computing (ICSTCC), pp. 633–638, IEEE, 2022. (WoS)
- T.-A Sauciuc, A. Burlacu, L. Ferariu, **A.-P. Botezatu**, "SO_3 - CNN: Learning rigid displacement using depth images and orthogonal dual tensors," in IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, pp. 1–6, IEEE, 2022.
- F.-A. Brașoveanu, A.-I. Iancu, **A.-P. Botezatu**, and A. Burlacu, "3D Vision-based Workspace with Deep Learning Capabilities for Autonomous Robot Manipulation," in New Advances in Mechanisms, Mechanical Transmissions and Robotics, pp. 244–253, Springer Nature Switzerland, 2025
- **A.-P. Botezatu** and A. Burlacu, "CNN-Based Early Fusion with Intensity Features for Road Surface Classification" (submitted to ICSTCC 2025)

7.2 Contribuții și direcții viitoare de cercetare

Contribuția principală a acestei teze constă în dezvoltarea unor arhitecturi inovatoare de învățare profundă cu fuziune timpurie și a unor sisteme hibride care îmbunătățesc clasificarea

suprafeței drumului și aplicațiile de control cu feedback vizual. Aceste inovații reprezintă progrese semnificative în domeniile respective, demonstrând îmbunătățiri față de metodele tradiționale și stabilind noi abordări pentru integrarea datelor multimodale în sistemele inteligente.

Pe baza cercetării prezentate în această teză, se conturează mai multe direcții promițătoare pentru lucrări viitoare. Arhitectura cu fuziune timpurie pentru clasificarea suprafețelor rutiere ar putea fi extinsă pentru a integra senzori suplimentari (LiDAR, radar, imagini termice) și informații temporale prin rețele neurale recurente sau transformere, îmbunătățind robustețea în condiții meteorologice extreme. Pentru servoing vizual, cercetarea viitoare ar putea explora reprezentări alternative în timp real (flux optic, hărți de adâncime) și dezvoltarea capacității de a prezice atât comenzi de viteză, cât și poza 3D a robotului. Arhitectura hibridă bazată pe SuperPoint ar beneficia de mecanisme de învățare conștiente de erori pentru a aborda incertitudinile în potrivirea punctelor cheie, iar integrarea tehnicilor de AI explicabilă ar putea oferi perspective valoroase în procesul de luare a deciziilor, îmbunătățind transparența sistemului și ghidând rafinamente arhitecturale ulterioare.

Bibliografie

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[2] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[3] F.-A. Brașoveanu, A.-I. Iancu, A.-P. Botezatu, and A. Burlacu, "3-d vision-based workspace with deep learning capabilities for autonomous robot manipulation," in *New Advances in Mechanisms, Mechanical Transmissions and Robotics*, pp. 244–253, Springer Nature Switzerland, 2025.

[4] E. Ribeiro, R. Mendes, and V. Grassi, "Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation," *Elsevier's Robotics and Autonomous Systems*, vol. 139, p. 103757, 2021.

[5] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[6] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[7] M. Nolte, N. Kister, and M. Maurer, "Assessment of deep convolutional neural networks for road surface classification," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 381–386, IEEE, 2018.

[8] T. Zhao, J. He, J. Lv, D. Min, and Y. Wei, "A comprehensive implementation of road surface classification for vehicle driving assistance: Dataset, models, and deployment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8361–8370, 2023.

[9] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

[12] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[13] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[14] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, 1992.

[15] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 d visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 2, pp. 238–250, 1999.

[16] A. Saxena, H. Pandya, G. Kumar, A. Gaud, and K. M. Krishna, "Exploring convolutional networks for end-to-end visual servoing," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3817–3823, IEEE, 2017.

[17] Q. Bateux, E. Marchand, J. Leitner, F. Chaumette, and P. Corke, "Training deep neural networks for visual servoing," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3307–3314, IEEE, 2018.

[18] Z. Machkour, D. Ortiz-Arroyo, and P. Durdevic, "Classical and deep learning based visual servoing systems: a survey on state of the art," *Journal of Intelligent & Robotic Systems*, vol. 104, no. 1, p. 11, 2022.

[19] A.-P. Botezatu and A. Burlacu, "A short review of deep learning methods in visual servoing systems," *Bulletin of the Polytechnic Institute of Iași. Electrical Engineering, Power Engineering, Electronics Section*, vol. 69, no. 3, pp. 113–136, 2023.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

[21] A. Ganciu, A. Burlacu, and R.-L. Roşca, "A pose-free 2D-3D hybrid approach for visual servoing control systems," in *2024 28th International Conference on System Theory, Control and Computing (ICSTCC)*, pp. 527–532, IEEE, 2024.