

Detectarea reperelor urbane folosind Vederea Artificială

Teză de doctorat – Rezumat

pentru obținerea titlului științific de doctor la

Universitatea Politehnică Timișoara

în domeniul de doctorat INGINERIE ELECTRONICĂ, TELECOMUNICAȚII ȘI
TEHNOLOGII INFORMAȚIONALE

autor ing. Ciprian-Constantin Orhei

conducător științific Prof.univ.dr.ing. Radu Vasii

luna 05 anul 2022

1 Motivație

Reperele sunt de obicei definite din două perspective: una ca un obiect sau o structură care este ușor de vizibil și de recunoscut, iar a doua ca o clădire sau un loc care are o importanță istorică importantă. Reperele dintr-o zonă urbană servesc drept „magnet spațial” în care au loc activități culturale, civice sau economice. În acest sens au devenit un aspect important în multiple domenii legate de turism și cultură [1][2].

Identificarea și localizarea unui reper urban formează o activitate care îmbină în mod natural mai multe domenii de cercetare precum procesarea semnalului de imagine (ISP), viziunea computerizată (CV), realitate augmentată (AR). Această îmbinare a mai multor domenii a fost primul imbold care m-a determinat să aleg această temă de cercetare pentru teză.

Interacțiunea umană dintre repere și ecosistemul orașelor a devenit un subiect de interes pentru mine datorită a două evenimente: unul este decernarea orașului meu natal, Timișoara, a titlului de Capitală Europeană a Culturii pentru anul 2023, respectiv, interacțiunilor mele cu proiectul Spotlight Heritage Timișoara.

Ca rezultat a acestei teze, doresc să ofer o soluție de detectare a reperelor urbane, din perspectiva street view, care poate fi utilizată într-o aplicație AR mobilă de turism. Această direcție dorește să se alinieze cu dezvoltarea continuă a aplicațiilor mobile destinate Timișoara Capitală Europeană a Culturii 2023. Rezultate a sistemului de detecție de repere propus sunt prezentate în Figura 1.1.

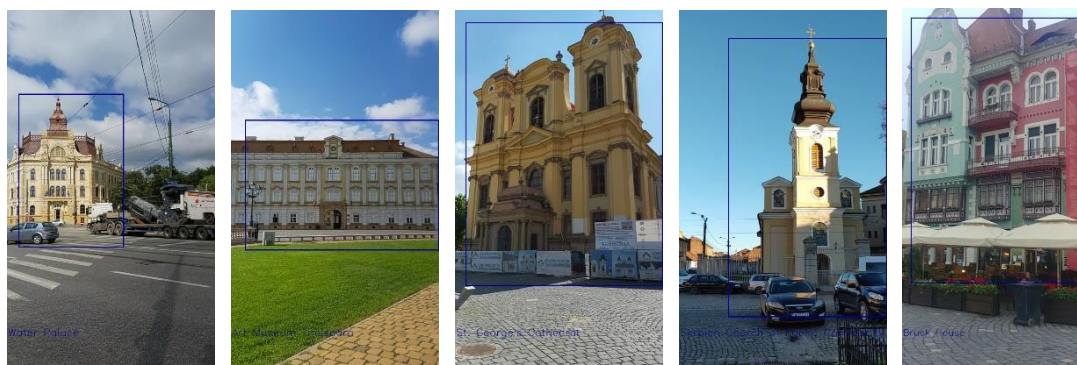


Figura 1.1 Exemplu de detectare a reperelor urbane propuse

În această teză voi încerca să răspund la următoarele întrebări de cercetare:

1. Care este “state of the art” în detectarea reperelor urbane folosind imagini capturate folosind camere mobile?
2. Ce ar trebui să ofere un framework de simulare pentru a fi considerat o soluție

potrivită pentru sistemele de procesare de această natură?

3. Ce algoritmi ISP îmbunătățesc imaginea pentru a obține o detecție mai bună în acest caz?

4. Care sunt provocările în crearea unei soluții de detectare a reperelor urbane adaptată pentru utilizare în mediul urban Timișorean?

Teza este structurată în mai multe capitole care sunt descrise mai jos.

Capitolul 1 este o expunere a motivației mele față de alegerea subiectului acestei teze. Prin scurta expunere, doresc să explic interconexiunile mai multor domenii care au fundamentat decizia de a alege această temă de cercetare.

Capitolul 2 oferă o privire de ansamblu asupra domeniului de detectare a reperelor urbane, de la aspectele generale focalizând spre final la un subdomeniu specific al sistemului de regăsire a imaginilor bazat pe conținut. Capitolul se concentrează pe prezentarea ecosistemului domeniului cu toate provocările și soluțiile pe care literatura le are de oferit.

Capitolul 3 își propune să prezinte sistemul de simulare ales de mine. Capacitatea de a simula offline un sistem este una esențială, cu beneficii considerabile în dezvoltare. End-to-End Computer Vision Framework (EECVF) [3][4] este un framework open source, bazat pe python, cu scopul de a oferi un instrument flexibil și dinamic pentru cercetare.

Capitolul 4 prezintă un algoritm propus de sharpening a imaginii, cu un consum de resurse scăzut și bazat pe filtre dilatate [5][6]. Algoritmul propus este evaluat pe mai multe cazuri de utilizare care pot apărea într-un sistem de detectare a reperelor pentru a înțelege mai bine beneficiile.

Capitolul 5 prezintă algoritmul de detectare a reperelor propus, cu o aprofundare în fiecare bloc constitutiv al acestuia. Am încercat ca fiecare decizie arhitecturală din interiorul algoritmului să fie explicată și justificată în contextul nostru de utilizare. Evaluarea algoritmului de detecție propus se face utilizând un set de date popular, prezentat în Capitolul 2, plus setul de date specific Timișoarei care a fost creat pentru acest scop.

Capitolul 6 este partea de concluzie a tezei, începând cu câteva concluzii generale referitoare la cercetările pe care am făcut-o. Ulterior, voi continua cu enumerarea contribuțiilor teoretice și practice pe care această teză le aduce în domeniile științifice.

2 Detectarea de repere urbane

Recunoașterea clădirilor sau a reperelor în mediile urbane urmărește să facă distincția între diferitele clase unice într-un set de date de imagine la scară largă. Acest amestec de tehnologii cu scopul de a recunoaște repere este folosit în alte câteva domenii în zilele noastre, cum ar fi jocurile pe computer, planificarea urbană, industria divertismentului, realizarea de filme sau cartografierea digitală pentru orientare. [7].



Figura 2.1 (a) Exemplu de imagini din seturile de date de recunoaștere a clădirii; (b) Exemplu de imagini din setul de date de înțelegere a mediului urban.

În general, recunoașterea reperelor este o sarcină provocatoare în domeniul CV-ului. Pot apărea mai multe provocări atunci când încercăm să potrivim mai multe imagini din același loc,

provocari cum ar fi modificări ale condițiilor de iluminare și ale punctului de vedere, sau prezența unor elemente de distracție, cum ar fi copaci, oameni sau semne. Pentru a atenua aceste provocări, abordările existente se bazează pe descrierea caracteristicilor cu un anumit grad de invarianță la scară, orientare și iluminare [8].

Pentru a înțelege mai bine provocările, le-am divizat în tematici mai mici pe care le-am tratat separat: seturi de date pentru recunoașterea reperelor, seturi de date pentru înțelegerea mediului urban, soluții de detectare și provocări specifice, evaluarea detecției.

Evaluarea comparativă a unui sistem de detectare a clădirilor nu este un lucru banal de făcut. Aceasta este o sarcină complexă din cauza varietăților de ambele părți, algoritmului de detectare și schema de referință. Construirea unui punct de referință general valabil pentru detecție este aproape imposibil din cauza indiciilor diferiți utilizați în procesul de detectare sau a unicității reperelor în cazul dat. Exemple sunt prezentate în Figura 2.1(a).

Perspectiva imaginilor care creează setul de date este importantă pentru multe sisteme de detectare. Unele scheme de detectare se bazează pe faptul că imaginile de intrare sunt dintr-o perspectivă „street-view”. Această ipoteză permite proiectanților sistemului să utilizeze elemente specifice sau decizii în lanțul de procesare.

Recunoașterea automată a obiectelor scenei urbane are ca scop clasificarea și segmentarea imaginii unui mediu urban în categorii precum clădiri, mașini, oameni, vegetație și așa mai departe. Exemple de imagini din seturile de date de înțelegere a mediului urban sunt prezentate în Figura 2.1(b). În zilele noastre, datorită evoluției algoritmului CV, înțelegerea scenei se face de obicei prin segmentare semantică.

Evaluarea bazată pe clasament sunt prezente în literatură și sunt din ce în ce mai populare. Într-un fel, dacă luăm în considerare că detectarea claselor (repere în cazul nostru) nu reprezintă probleme pură bazate pe imagini, metrica de clasare este mai adecvată. O metrică care se referă la evidențierea succesului pozitiv al detectării este *TopK*, prezentată în ecuația (2.1) [9].

$$TopK = \frac{1}{|Q|} \sum_{q \in Q} \text{Imagini relevante recunoscute în primele } K \text{ imagini} \quad (2.1)$$

În Figura 2.2 prezint schema generică pentru detecția de repere. După cum putem observa, sistemul va include întotdeauna două faze: faza offline, în care pregătim vectorii de caracteristici pentru antrenarea clasificatorului sau a modelului de clasificare în cazul soluțiilor bazate pe rețele neuronale convoluționale (CNN) și faza online în care folosim o imagine pentru a interoga clasificatorul în vederea obținerii unei detecții. Unele dintre blocurile detaliate în Figura 2.2 sunt opționale (marcate cu linii întrerupte) și sunt decizii care sunt luate de proiectarea sistemului.

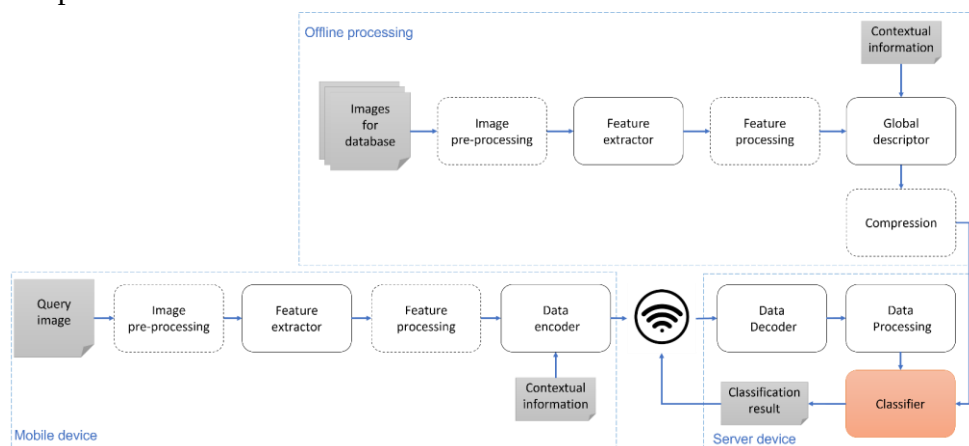


Figura 2.2 Schemă generică pentru detectarea reperelor folosind dispozitive mobile

În general, vedem că soluțiile propuse în literatură variază din punct de vedere al caracteristicilor folosite până la soluțiile de grupare și potrivire. În ultimii ani putem observa o tendință clară bazată pe CNN în soluțiile propuse. Din punctul de vedere al caracteristicilor utilizate, gama de soluții propuse în literatură variază de la caracteristici clasice, așa-numite de nivel inferior (cum ar fi muchile, joncțiunile) la caracteristici globale sau locale până la soluții bazate pe CNN.

Prima dificultate reală apare în acuratețea locației GPS a datelor care pot îndepărta în zonele dense. Proiectarea sistemelor de recuperare ar trebui să gestioneze și să compenseze sistemul de etichete GPS predispus la erori, chiar dacă geo-etichetarea a progresat considerabil în ultimii ani.

A doua provocare este informația care este surprinsă de imagini. În scenariile urbane, cantitatea de distracție este considerabilă, iar în acest domeniu sunt deja propuse soluții multiple pentru sistemele mobile de recunoaștere a reperelor.

A treia problemă pe care trebuie să o abordeze sistemele este compromisul dintre acuratețea recunoașterii și constrângerile cerințelor în timp real. La provocarea timpului de rulare și a resurselor se adaugă limitarea lățimii de bandă a dispozitivelor mobile actuale din rețelele de telecomunicații.

3 End-To-End Computer Vision Framework

Arhitecturile de CV actuale includ pași multipli pornind de la achiziția de imagini de la senzori, procesare pentru îmbunătățirea imaginii, transformări în vederea reducerii zgomotului din imagini, selectarea regiunii de interes, nivele diferite de extracție de caracteristici, procesare relevantă abstractă pentru aplicații și luarea deciziilor, cum ar fi clasificarea unui obiect [10].

Pentru a putea simula sistemul propus, am folosit framework-ul CV bazat pe Python, EECVF [4]. Motivația utilizării acestui cadru s-a bazat pe mai multe elemente: versatilitatea limbajului de programare utilizat, capacitatea de extindere a algoritmilor CV existenți, capacitatea de expunere a informațiilor de depanare. Desigur, familiaritatea cadrului EECVF a fost un factor important, dar nu singular.

Un alt aspect interesant pe care îl pot sublinia este limbajul principal de programare folosit în dezvoltarea framework-urilor. Putem vedea clar o corelație între nivelul de maturitate al limbajului Python și utilizarea acestuia în construirea framework-urilor. Acest aspect este firesc, din punctul meu de vedere, deoarece Python oferă funcții build-in pentru a interconecta mai multe servicii furnizate de sisteme de operare sau acceleratoare HW.

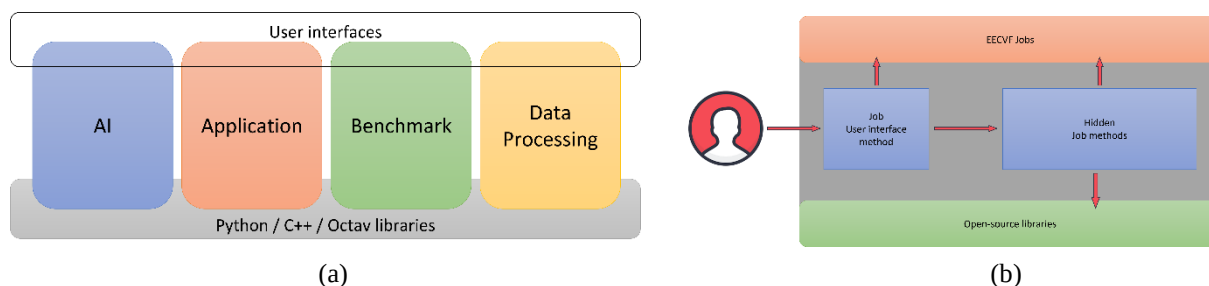


Figura 3.1 (a) Blocurile constructoare a EECVF; (b) Structura de ansamblu a unui Job

Framework-ul poate gestiona ciclul de creare - antrenament - evaluare a unui model AI, executarea unei aplicații CV folosind modelul și, în final, evaluarea rezultatelor și afișarea lor fără intervenția utilizatorului. De la capacitatea aceasta de a gestiona acest lanț, numele de „End-to-End”.

EECVF este un framework ușor de utilizat, modular și flexibil, conceput pentru cercetarea și testarea conceptelor CV. Prezentarea generală blocurilor este prezentată în Figura 3.1 (a). Framework-ul nu necesită ca utilizatorul să gestioneze interconexiunile în cadrul sistemului [3]. Desigur, acest lucru nu este specific acestui framework, dar modul trivial în care noi algoritmi de CV pot fi conectați în EECVF este un aspect important în decizie.

Fiecare job are o interfață publică, prezentă în blocul User Interface, din care poate fi configurată prin parametri. Această metodă publică ar trebui să gestioneze toate modificările necesare în interiorul blocurilor și chiar să declanșeze alte joburi, din același sau bloc diferit, dacă este necesar. În Figura 3.1(b) este prezentată structura internă a unui job.

Un dezavantaj al EECVF-ului este lipsa unui GUI pentru utilizatorii începători, dar în cazul nostru de utilizare nu este cazul. În plus, interesul meu a fost mai mult în dezvoltarea și extinderea de noi algoritmi de CV în framework decât în facilitarea folosirii celor existente.

4 Filtre dilate

Prima mențiune despre dilatare în domeniul procesării imaginii apare în domeniul morfologiei matematice. Operația de dilatare folosește un element de filtru pentru a sonda și extinde formele conținute într-o imagine [11].

A doua mențiune a ideii de filtre dilatate a apărut în domeniul ML în ultimii ani. Este o tehnică în care se extind filtrele prin introducerea unor găuri între elementele consecutive [12].

A treia mențiune a filtrelor dilatate în literatură este ideea similară din ML, dar aplicată la convoluțiile clasice bazate pe filtre în afara oricărui model AI. Experimentele au fost efectuate pe filtre de detectare a muchiilor cu rezultate bune [4], [5], [13].

Pentru a beneficia de o vecinătate mai mare de pixeli pentru a obține o muchie, definim filtrul dilatat ca extinderea filtrului original cu un factor de dilatare. Când filtrele sunt dilatate, pozițiile nou adăugate sunt considerate ca lacune și le ignorăm prin completarea cu zerouri [5]. Obiectivul principal al dilatării este de a acoperi mai multă informație de la intrare în procesarea operației de convoluție. Prin aplicarea acestei operații, rezultă un câmp vizual mai larg la același cost computațional.

Rezultate pozitive au fost obținute prin utilizarea filtrelor dilatate în cadrul mai multor algoritmi de detectare a muchiilor, lucru care a fost efectuat în [4] [5]. În Figura 4.4 sunt prezentate rezultatele vizuale când se utilizează unul dintre cei mai populari algoritmi clasici de detectare a muchiilor: Canny [14]. După cum putem observa în imagini, fiecare nivel de dilatare poate aduce cu el puncte de muchie suplimentare

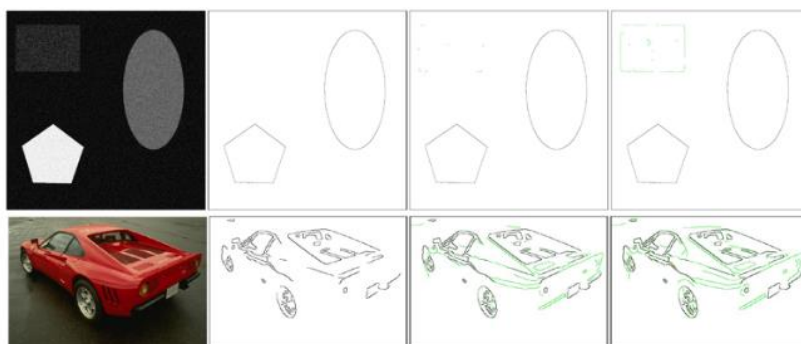


Figura 4.1 Harta muchiilor rezultate folosind algoritmul Canny cu aceiași parametri. Coloanele reprezintă imaginea originală, harta muchiilor folosind filtrul 3x3, folosind filtrul dilatat 5x5, folosind filtrul dilatat 7x7. Sursa imaginii este [9].

Încurajat de rezultatele pozitive pe care filtrele dilatate le aduc procesului de detectare

a muchiiilor, acum aş dori să investighez modul în care această dilatare afectează procesul de ascuţire.

Îmbunătăţirea contrastului imaginii este o preocupare legată de ascuţire a anumitor caracteristici, cum ar fi muchiile, limitele obiectelor sau texturile. Scopul principal al acestei activităţi este de a îmbunătăţi aspectul vizual al imaginii

Au fost avansate diferite abordări pentru îmbunătăţirea contrastului. Una dintre cele mai frecvente metode utilizate este egalizarea histogramei. O altă soluţie populară pentru ascuţire este filtrul trecere-sus (HP). Filtrele precum trece-jos, respingerea benzii sau HP sunt identificate ca fiind filtre ideale. Un filtru ideal are proprietatea că reduce toate frecvenţele peste (sau sub) cu un anumit prag de frecvenţă [15].

O altă soluţie populară pentru îmbunătăţirea imaginii este tehnica de mascare neclară (UM), această tehnică a apărut în fotografie cu scopul de a îmbunătăţi calitatea imaginilor prin mascarea detaliilor acestora. În această variantă am avea în vedere scăderea unei copii neclare a imaginii din imaginea originală în sine. Chiar dacă soluţia este una simplă, aceasta are câteva dezavantaje, cum ar fi sensibilitatea ridicată la zgomot şi posibile îmbunătăţiri ale zonelor cu contrast ridicat [16].

În acest scop am dilata următoarele filtre Laplace găsite în literatură [15]. Făcând acest lucru, am dori să luăm în considerare atunci când aplicăm ascuţirea unui vecinătăţi mai largi. Pentru a evalua corect dacă acesta este un beneficiu, trebuie să luăm în considerare şi filtrele extinse 5x5.

$$\begin{array}{cccccc}
 \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} &
 \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 1 & 2 & -17 & 2 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} &
 \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -4 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} &
 \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} &
 \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -18 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} &
 \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -8 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \\
 \text{V1 } 3 \times 3 & \text{V1 } 5 \times 5 & \text{V1 } 5 \times 5 \text{ dilated} & \text{V2 } 3 \times 3 & \text{V2 } 5 \times 5 & \text{V2 } 5 \times 5 \text{ dilated}
 \end{array}$$

Figura 4.2 Filtre Laplace: 3x3, 5x5 şi 5x5 dilatate

Evaluarea unei imaginii îmbunătăţite este o sarcină dificilă din două motive: nu există un indice specific care să determine nivelul optim de claritate şi faptul că orice algoritm de claritate pe care l-am folosi are mai mulţi parametri care pot modifica rezultatul. Din literatura de specialitate am ajuns la concluzia că următoarele metrici sunt utilizate pentru a evalua şi a judeca efectul ascuţirii: entropia unei imaginii, factorul spaţial şi contrastul mediu [17] [18].

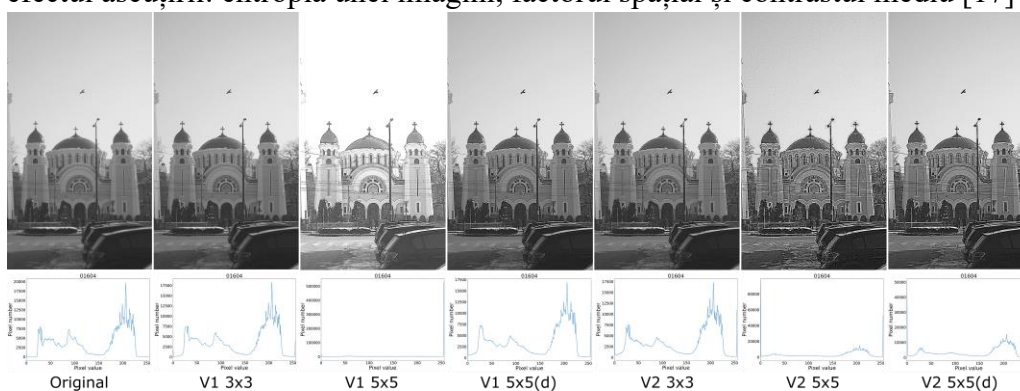


Figura 4.2 Rezultate vizuale ale HP folosind Laplace V1 şi V2

În Figura 4.3 ni se prezintă un exemplu vizual al rezultatelor obţinute pe imaginea „01604” din setul de date TMBuD [19]. După cum putem observa la utilizarea filtrelor dilatate, imaginile sunt îmbunătăţite mai bine decât cu 3x3 clasic, dar mai mici decât atunci când se utilizează 5x5. Din histograme putem concluziona că utilizarea filtrele V1 nu aplatizează atât de mult histograma ca în cazul filtrelor V2.

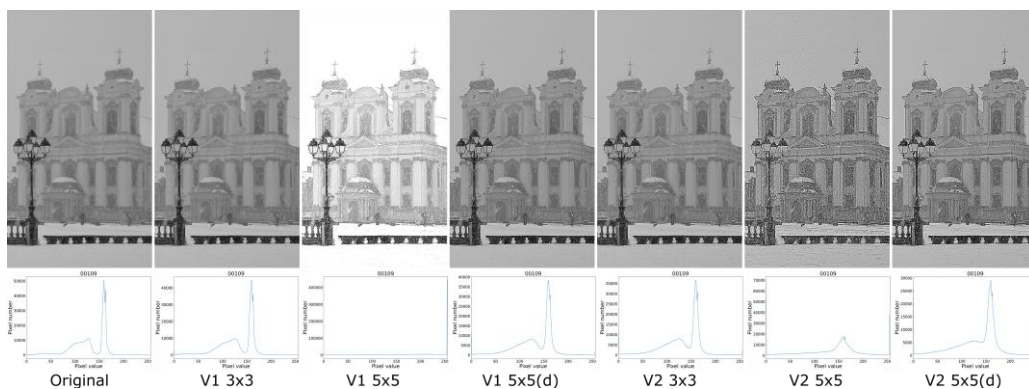


Figura 4.3 Rezultate vizuale ale UM folosind filtre Laplace V1 și V2

În Figura 4.4 sunt prezentate rezultatele vizuale pentru imaginea „00109” din setul de date TMBuD [19]. Ca și înainte, putem observa că utilizarea filtrului dilatat 5x5 aduce rezultate mai bune decât 3x3 clasic. Un alt aspect demn de menționat este că folosirea filtrului extins 5x5 nu pare a fi o soluție bună pentru această activitate.

Tabelul 4.1 Rezultate medii pe întregul set de date pentru HP și UM

	HP		UM	
	Entropy	SF	Entropy	SF
Original image	7.105	14.881	7.105	14.881
HP V1 3x3	7.234	31.383	7.234	31.383
HP V1 5x5	5.143	69.286	5.143	69.286
HP V1 5x5 dilated	7.236	43.032	7.236	43.032
HP V2 3x3	7.226	49.896	7.226	49.896
HP V2 5x5	6.396	92.332	6.396	92.332
HP V2 5x5 dilated	6.966	62.801	6.966	62.801

Din Tabelul 4.1, am concluzionat că utilizarea filtrelor dilatate pentru clarificarea imaginii este o cale care merită urmată. În scopul nostru, am analizat doar fezabilitatea conceptului folosind soluții clasice standard de ascuțire. Această limitare a domeniului de aplicare este determinată și de un factor practic, algoritmul de detectare a reperelor propus este nevoie de o soluție de calcul rapidă.

5 Sistem de detectare a reperelor urbane propus

Recunoașterea reperelor (clădirii) este considerată a fi o problemă de detectare a obiectelor sau de regăsire a imaginilor bazată pe conținut pentru un anumit domeniu de aplicare. În comparație cu sarcina generală de recunoaștere a obiectelor, aceasta specifică aduce mai multe provocări, deoarece majoritatea imaginilor urbane conțin atât obiecte create de om, cât și obiecte naturale.

Imaginile făcute cu aceeași clădire ar putea demonstra o gamă largă de variabilitate - pot fi luate din puncte de vedere diferite, în condiții de iluminare diferite sau pot suferi de ocluzii parțiale de la copaci, vehicule în mișcare, alte clădiri. Prin urmare, o tehnică ideală de recunoaștere a clădirii ar trebui să fie suficient de sensibilă pentru a identifica o clădire individuală, în aceeași timp robustă la diferite transformări geometrice și fotometrice de imagine [7].

Algoritmul propus este capabil să gestioneze situații sau scenarii dificile când mai multe repere de interes sunt grupate într-un mic areal, cum ar fi piețele orașului sau zone

turistice. Aceste situații sunt cele mai interesante din perspectiva aplicațiilor, deoarece aplicațiile turistice au de obicei dificultăți în acest tip de situații.

5.1 TMBuD – set de date pentru detectare

Setul de date TMBuD de detectare a clădirii este o extensie a setului de date TMBuD [19]. Acest set de date își propune să susțină o evaluare cantitativă a căutării vizuale mobile susținute de context GPS, a unui subset de repere din zona Timișoarei. Principala noutate a acestui set de date este scopul clar vizat de a conține repere doar din orașul Timișoara, care este principalul areal vizat de sistemul nostru de detectare a reperelor.

Setul de date de clădire TMBuD conține 1097 de imagini cu rezoluția de 768x1024 pixeli, capturate folosind un telefon mobil din perspectiva street-view. Imaginile sunt organizate în 125 de repere raionale situate în mai multe zone turistice din Timișoarei



Figura 5.1 Exemplu de reper unic în setul de date TMBuD

Un aspect important de specificat este că benchmark-ul cuprinde fotografiile capturate cu telefonul mobil ale reperelor urbane care urmăresc să includă variații în calitate, estompare, modificări de iluminare, obstrucții și diverse unghiuri de vizualizare. Acest aspect este prezentat în Figura 5.1 unde au fost selectate mai multe imagini dintr-un singur reper. Intenția a fost să capteze fotografiile de la distanțe diferite (aproape, mijlocie, lungă) și diferite unghiuri de direcție pentru a putea imita scenariul din viața reală.

Setul de date de referință propus are perspective și condiții diferite, așa cum sa menționat anterior, dar oferă posibilitatea de a evalua un sistem în două condiții unice: imagini de rezoluție scăzută capturate și condiții de noapte. După cum este prezentat în Figura 5.2, setul de date oferă un număr limitat de imagini de la repere la o rezoluție scăzută, mai exact 7,66% dintre ele, și imagini luate în timpul nopții, 14,9% dintre ele.

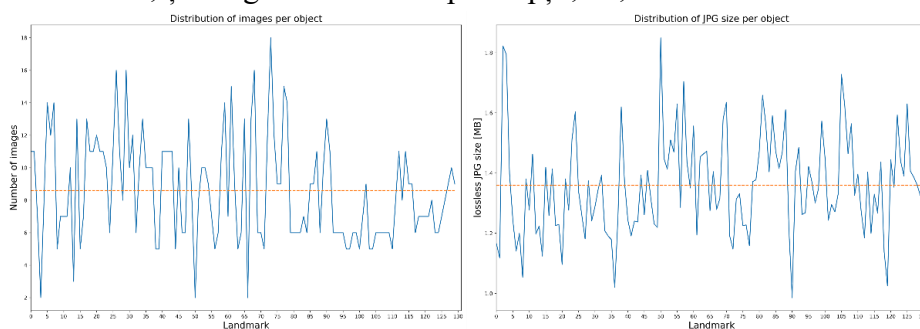


Figure 5.1 Number of images (left) and lossless JPG size (right) for each landmark from the dataset

Pentru a măsura diversitatea fotografiilor pentru fiecare reper, folosim dimensiunea medie de compresie JPEG (fără pierderi) a imaginilor. Această măsură se face prin calculul

imaginii medii a fiecărui reper și crearea fișierului JPG fără pierderi. Mărimea fișierului dezvăluie cantitatea de informații disponibile. Teoria din spatele acestui lucru este că o imagine diversă va avea ca rezultat o medie mai neclară, în comparație cu una care are puțină diversitate va avea ca rezultat una mai structurată, mai clară. Prin urmare, ne putem aștepta să vedem o dimensiune mai mică a fișierului JPG a imaginii medii pentru un reper care oferă imagini mai diverse. Distribuția în setul de date poate fi observată în Figura 5.2.

5.2 Sistemul de detecție propus

În Figura 5.3 se prezintă arhitectura generală de procesare propusă pentru cazul de utilizare. Fluxul propus este prezentat în detaliu în următorul subcapitol și evaluat ulterior.

Dacă dorim să rezumăm toți pașii necesari acestei acțiuni, putem enumera următoarele: (i) colectarea imaginilor și a metadatelor; (ii) preprocesarea imaginii (sarcina necesară pentru a îmbunătăți clasificarea finală); (iii) extractor de caracteristici; (iv) generarea cărții de coduri; (v) clasificator.

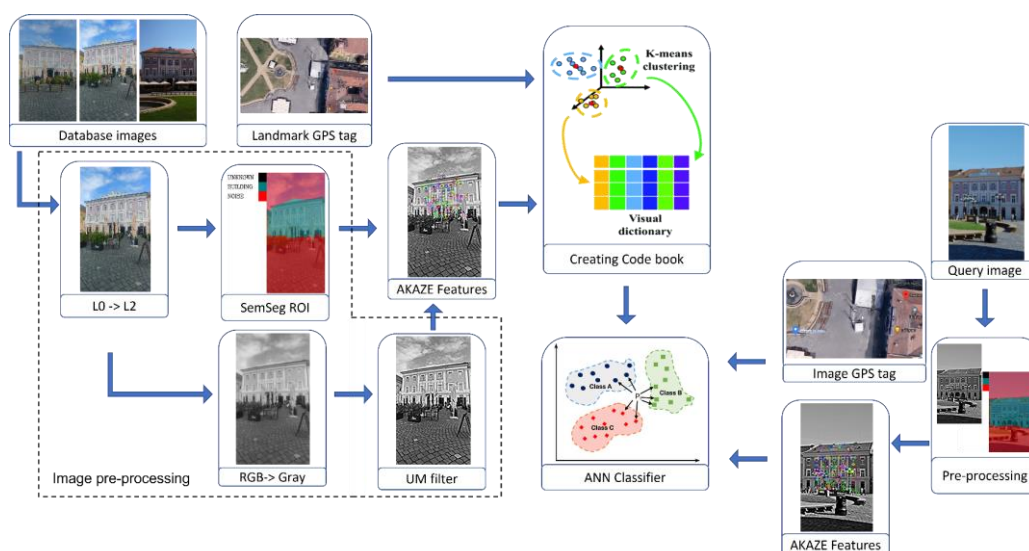


Figura 5.3 Flux propusă pentru detectarea reperelor

În calea offline a sistemului propus acesta preia imaginile setului de date și generează un vector de caracteristici A-KAZE [20] care este filtrat folosind o regiune de interes. Această acțiune în comparație cu alte soluții nu adaugă o pondere mai mică caracteristicilor din afara regiunii, ci le ignoră pe toate împreună. Presupunerea este că acestea sunt generate din distractorii din imagine. Ulterior, vectorii de caracteristici sunt grupați împreună într-un vector pentru fiecare reper, care sunt grupați într-o structura BOF [21] folosind algoritmul de clustering KNN [22].

În faza on-line oglindim partea de pre-procesare pentru a filtra caracteristicile care sunt generate de distractoare și folosind clasificatorul ANN [23] găsim vectorul de reper similar cel mai apropiat de imaginea solicitată.

Pentru a îmbunătăți detectarea, folosim etichete GPS pentru a ne limita intervalul de căutare în cadrul clusterelor. Procedând astfel, obținem beneficii în direcția acurateței detectării și a timpului de procesare.

Pentru a genera caracteristici mai bune, în ceea ce privește localizarea și cantitatea, îmbunătățim imaginile folosind algoritmul UM de ascuțire dilatată. Configurația UM este aleasă cu grijă, astfel încât imaginile de calitate scăzută să fie îmbunătățite, dar imaginile de

bună calitate să nu devină prea „arse”.

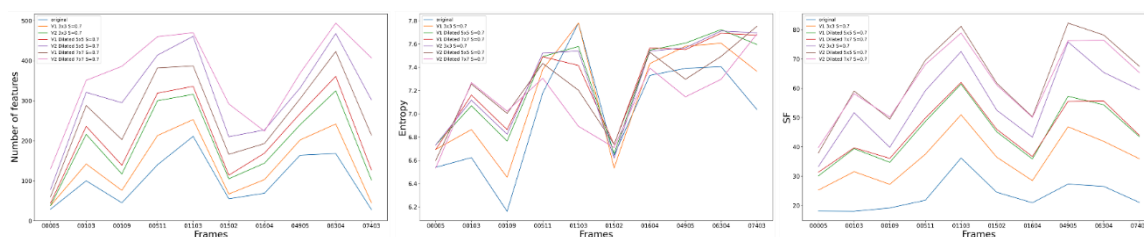


Figure 5.2 Number of features and image proprieties for pyramid level 2

În Figura 5.4, ne sunt prezentate rezultatele pentru întregul set de date utilizat. Putem observa că în fiecare caz clarificarea imaginii a dus la un număr mai mare de caracteristici detectate. Mai mult, la utilizarea filtrelor dilatate, numărul a crescut substanțial. Din datele statice prezentate, pot concluda că beneficiile clarificării imaginii folosind filtru dilatat înainte de etapa de extracție a caracteristicilor aduce beneficii, indiferent de nivelul piramidei care este procesat.

Pentru sistemul nostru am investigat o metodă bazată pe DL pentru alegerea regiunii noastre de interes, așa că vom urmări să realizăm o segmentare semantică clasică. Pentru sistemul nostru am ales Residual Networks cu 50 de straturi (ResNet-50) ca model de bază [24] al rețelei noastre cu SegNet [25] ca model de segmentare.

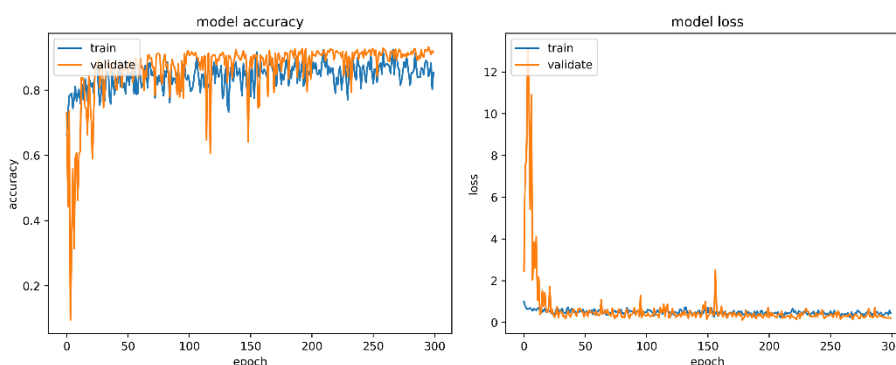


Figure 5.5 ResNet50-SegNet training model metrics

În Figura 5.5 prezint valorile reprezentate pentru Accuracy și Loss pentru procesul de antrenare. Cele două grafice arată integritatea și corectitudinea procesului de antrenament. Primul grafic urmărește precizia și oferă o perspectivă valoroasă asupra cantității de supra-adaptare care are loc în model. Dacă diferența dintre graficul de instruire și validare este mare, modelul este puternic supra-adaptat. Al doilea grafic urmărește pierderea din antrenament, deoarece evaluarea loturilor individuale se face în timpul trecerii propagate. Pierderea poate deveni mai liniară cu o rată scăzută de învățare.

5.3 Evaluarea sistemului de detecție propus

Pentru această secțiune, voi folosi următoarele notații: dimensiunea descriptorului A-KAZE (D_S), numărul de octave A-KAZE (D_NO), numărul de straturi A-KAZE (D_NL), pragul A-KAZE (D_THR), dilatăta dimensiunea filtrului pentru UM (UM_D), puterea UM a netezirii (UM_S), dimensiunea clusterului BOF (B_S), pragul pentru distanța pentru potrivirea caracteristicilor (THR), distanța în m între etichetele GPS (D_GPS).

Ca măsurătoare pentru evaluarea noastră, am ales să folosesc metrica *Top1*, prezentată în ecuația (1). Dacă luăm în considerare sistemul de detectare a reperelor, mai mult decât CBIR, această metrică are mai mult sens decât *mAP*.

5.3.1 Evaluare pe set de date publice

În primul rând, aş dori să evaluez performanţele algoritmului propus pe seturi de date publice pentru a avea o comparaţie cu alţi algoritmi existent. Din analiza făcută în capitolul 2.2, privind seturile de date şi rezumatul făcut în Tabelul 2.4, aleg să evaluez sistemul propus pe două seturi de date populare: ZuBuD [26] şi ZuBuD+ [8].

Pentru acest experiment, am scalat toate imaginile ZuBuD la dimensiunea 240x320. Această transformare a imaginilor de intrare a fost realizată prin alte soluţii găsite în literatură. Scopul acestei scalari este de a uşura procesarea.

Tabelul 5.1 Top1 rezultate pe setul de date

Soluţie	Configuraţie	ZuBuD	ZuBuD+
[8]	-	100.00%	100.00%
Propus fără ROI pe L0	D_S=8, D_NO=5, D_NL=6, D_THR=0.0012, UM_D=7, UM_S=0.1, B_S=300, THR=0.82	99.13%	99.80%
[27]	-	99.00%	99.00%
Propus cu L0	D_S=8, D_NO=5, D_NL=6, D_THR=0.001, UM_D=7, UM_S=0.5, B_S=150, THR=0.80	98.26%	99.60%
[28]	-	98.00%	-
[29]	-	95.00%	-
[30]	-	94.00%	-
Propus cu L1	D_S=8, D_NO=4, D_NL=6, D_THR=0.001, UM_D=7, UM_S=0.5, B_S=45, THR=0.80	93.91%	98.60%
Propus fără ROI on L1	D_S=8, D_NO=4, D_NL=6, D_THR=0.001, UM_D=5, UM_S=0.9, B_S=30, THR=0.82	92.17%	98.20%

După cum observăm în Tabelul 5.1, obţinem rezultate bune pe seturile de date ZuBuD şi ZuBuD+ cu algoritmul propus. Pentru această evaluare a sistemului propus, procesarea etichetelor GPS este dezactivată deoarece setul de date utilizat nu are această informaţie. Dar am experimentat cu diferite configuraţii ale nivelului piramidei şi ale selecţiei ROI.

5.3.2 Evaluare pe set de date propriu

În această secţiune vreau să evaluez algoritmul propus pe noul set de date TMBuD, care este prezentat în capitolul 5.2. Pentru a înţelege mai bine, folosesc toate cele trei configuraţii disponibile: Dataset 3_2 (3 vizualizări pentru antrenament şi 2 pentru evaluare), Dataset 3_5_Night (3 vizualizări pentru antrenament şi 5 pentru evaluare care conţin imagini cu scenarii nocturne), Dataset 3_N (număr dezechilibrat de imagini de testare).

În Tabelul 5.2, prezint toate configuraţiile şi metricile Top1 obţinute la evaluarea setului de date propus pe diferite variante ale acestuia. Ca primă concluzie preliminară, pot observa că la introducerea etichetei GPS rezultatele sunt îmbunătăţite pentru orice nivel de piramidă. A doua concluzie preliminară este faptul că utilizarea nivelului piramidal L3 nu obţine rezultate bune în general, chiar dacă consumul de resurse este redus dramatic. Lipsa rezultatelor pe L3 nu este ceva la care nu ne aşteptam, deoarece imaginea de intrare are o dimensiune de doar 90x160 pixeli.

Tabelul 5.2 Rezumatul experimentelor efectuate cu rezultatele Top1

Configuration	Lvl	GPS	3_2	3_5_Night	3_N
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=400, THR=0.8, D_GPS=100	L1	Yes	93.62%	93.59%	92.05%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=400, THR=0.8, D_GPS=100	L0	Yes	91.91%	91.81%	91.91%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=350, THR=0.8, D_GPS=100	L2	Yes	92.82%	92.17%	90.10%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=400, THR=0.8	L1	No	88.44%	87.90%	85.65%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=400, THR=0.8	L0	No	82.86%	87.90%	84.66%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=350, THR=0.8	L2	No	86.85%	83.27%	79.78%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=30, THR=0.8, D_GPS=75	L3	Yes	74.90%	72.60%	69.59%
D_S=8, D_NO=6, D_NL=3, D_THR=0.001, UM_D=7, UM_S=0.9, B_S=30, THR=0.8	L3	No	48.61%	43.77%	41.42%

În Figura 5.33, prezint analiza timpului de rulare pentru fiecare configurație evaluată în Tabelul 5.5. Când vorbim despre un sistem de detecție, de cele mai multe ori decizia vine făcută la un compromis între resurse și performanță. Primul lucru pe care îl putem observa este faptul că pentru imagini cu rezoluție mai mare, cum ar fi L0 și L1, filtrarea etichetelor GPS are un efect vizibil asupra duratei generale de rulare online. Dacă ne uităm la L2, care ar fi ținta noastră, deoarece considerăm un echilibru bun între performanță și resurse. Un alt aspect pe care trebuie să-l menționăm este că doar prin filtrarea utilizând GPS am redus timpul de execuție a interogării BOF cu 70%



Figure 5.6 Runtime analysis of proposed algorithm on different levels and configurations

Dacă ne uităm la durata de procesare online, putem observa că sistemul poate rula în jur de 6 cadre pe secundă (timpul total de rulare este de 160 ms fără job de achiziție a imaginii). Am eliminat jobul care obține imaginea deoarece într-un sistem în timp real procesul este diferit de cel pe care l-am folosit în experimentul nostru. Dacă luăm în considerare că în această etapă a dezvoltării algoritmul nu este optimizat pentru condiții de timp real, este o rată de cadre decentă la care se poate procesa.

Din Tabelul 5.1 și Tabelul 5.2, putem afirma clar că algoritmul propus are rezultate bune obținând o valoare de 99,13% Top1 pe setul de date ZuBuD și 92,05% pe setul de date TMBuD v3_N.

6 Contribuții și concluzii

Cercetarea de fond pentru această teză a fost realizată în cadrul Centrului de Cercetare Multimedia din cadrul Facultății de Electronică, Telecomunicații și Tehnologii Informaționale a Universității Politehnica din Timișoara. Activitatea de cercetare a fost completată de activități de voluntariat și a contribuit la aplicații mobile AR care deservește orașul Timișoara pentru titlul de Capitală Culturală Europeană în 2023.

Teza mea poate fi rezumată ca o propunere pentru o schemă de detectare a reperelor adaptată pentru mediului urban din Timișoara. Acest algoritm complex poate fi integrat într-o aplicație mobilă care poate oferi turiștilor șansa de a descoperi mai bine scenariul urban al orașului nostru.

În continuare, am încercat să răspund la întrebările de cercetare pe care le-am ridicat la începutul tezei, pe baza constatărilor cercetării mele:

1. Care este “state of the art” în detectarea reperelor urbane folosind imagini capturate folosind camere mobile?

O analiză și o trecere în revistă amplă a domeniului de detectare a reperelor este oferită în capitolul 2. Pentru a putea lua în considerare toate dificultățile pe care le putem întâmpina, problema a fost împărțită în părți mai mici: seturi de date repere și provocări, seturile de date de înțelegere a mediului urban și detectarea reperelor.

2. Ce ar trebui să ofere un framework de simulare pentru a fi considerat o soluție potrivită pentru sistemele de procesare de această natură?

O trecere în revistă a soluțiilor de framework existente este oferită în capitolul 3, care se concentrează pe evidențierea punctelor principale ale fiecărui cadru și în ce limbaj de programare au fost dezvoltate. Ca criteriu de alegere a unui cadru am considerat următoarele: (1) limbajul principal de programare utilizat; (2) Capacitatea de a simula un întreg lanț de prelucrare; (3) Capacitatea de a emite date intermediare și de depanare; (4) Integrarea ușoară a noilor algoritmi CV.

3. Ce algoritmi ISP îmbunătățesc imaginea pentru a obține o detecție mai bună în acest caz?

Pentru a răspunde la această întrebare, am apelat la conceptul de filtre dilatate, care sunt descrise în capitolul 4. Din literatură am descoperit că utilizarea filtrelor dilatate poate îmbunătăți conducta de procesare a imaginii. Acest concept a fost validat de la detectarea marginilor bazată pe filtru până la segmentarea semantică CNN. În acest sens, am aplicat cu succes conceptul dilatat la algoritmi de clarificare a imaginii și am obținut rezultate mai bune cu aceleași resurse de procesare necesare.

4. Care sunt provocările în crearea unei soluții de detectare a reperelor urbane adaptată pentru utilizare în mediul urban Timișorean?

Aceasta este întrebarea fundamentală la care sa răspuns în această teză și este abordată în capitolul 5. După cele menționate, aceasta este o problemă în două părți: setul de date specific necesar pentru aceasta și sistemul de recunoaștere a reperelor adaptat. Pornind de la informațiile adunate și structurate în capitolul 2, am putut oferi un set de date Timișoara de repere care este ușor scalabil în viitor și un sistem de recunoaștere a reperelor care poate fi ușor adaptat pentru o aplicație mobilă.

În teză am următoarele contribuții teoretice:

1. Prezentare generală a framework-urilor software existente pentru CV.
2. Analiza filtrelor dilatate în mai multe subiecte de CV.
3. Analiza algoritmului de detectare a muchiilor axat pe scenarii urbane.
4. Prezentare de ansamblu asupra seturilor de date de recunoaștere a clădirilor

(reper).

5. Prezentare de ansamblu asupra setului de date pentru înțelegerea mediului urban.
6. Prezentare generală a soluțiilor de recunoaștere a reperelor.
7. Baze teoretice pentru detecția urbană Timișoara.

În teză am următoarele contribuții teoretice:

1. Dezvoltarea framework-ului EECVF.
2. Dezvoltarea algoritmi de detecție de muchii clasici.
3. Aplicarea dilatării filtrelor în cadrul Filtru trece sus și Unsharp Masking.
4. Implementarea în EECVF a experimentelor realizate în teză.
5. Setul de date de recunoaștere TMBuD.
6. Training pentru segmentare semantică și evaluare pentru mediul Timișoara.
7. Sistem de recunoaștere a reperelor pentru mediul urban Timișoara.

BIBLIOGRAPHY

- [1] S. Vert, D. Andone, A. Ternauciuc, V. Mihaescu, O. Rotaru, M. Mocofan, C. Orhei, and R. Vasii, "User Evaluation of a Multi-Platform Digital Storytelling Concept for Cultural Heritage," *Mathematics*, vol. 9, no. 21, Art. no. 21, Jan. 2021, doi: 10.3390/math9212678, WOS:000750692000001 (Q1 journal)
- [2] K. Lynch, *The Image of the City*. MIT Press, 1964
- [3] C. Orhei, M. Mocofan, S. Vert, and R. Vasii, "End-to-End Computer Vision Framework," in 2020 International Symposium on Electronics and Telecommunications (ISETC), Nov. 2020, pp. 1–4. doi: 10.1109/ISETC50328.2020.9301078, WOS:000612681000017.
- [4] C. Orhei, S. Vert, M. Mocofan, and R. Vasii, "End-To-End Computer Vision Framework: An Open-Source Platform for Research and Education," *Sensors*, vol. 21, no. 11, Art. no. 11, Jan. 2021, doi: 10.3390/s21113691, WOS:000660684600001 (Q1 journal).
- [5] V. Bogdan, C. Bonchis, and C. Orhei, Custom Dilated Edge Detection Filters. Václav Skala - UNION Agency, 2020. doi: 10.24132/CSRN.2020.3001.19.
- [6] C. Orhei, V. Bogdan, C. Bonchis, and R. Vasii, "Dilated Filters for Edge-Detection Algorithms," *Appl. Sci.*, vol. 11, no. 22, 2021, doi: 10.3390/app112210716, WOS:000725536600001 (Q2 journal).
- [7] J. Li, W. Huang, L. Shao, and N. Allinson, "Building recognition in urban environments: A survey of state-of-the-art and future challenges," *Inf. Sci.*, vol. 277, pp. 406–420, Sep. 2014, doi: 10.1016/j.ins.2014.02.112.
- [8] F. Magliani, N. M. Bidgoli, and A. Prati, "A location-aware embedding technique for accurate landmark recognition," in *Proceedings of the 11th International Conference on Distributed Smart Cameras*, New York, NY, USA, Sep. 2017, pp. 9–14. doi: 10.1145/3131885.3131905.
- [9] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 593–601, Apr. 2001, doi: 10.1016/S0167-8655(00)00118-5.
- [10] R. Klette, *Concise Computer Vision*. London: Springer London, 2014. doi: 10.1007/978-1-4471-6320-6.
- [11] J. S. J. Lee, R. M. Haralick, and L. G. Shapiro, "Morphologic Edge Detection," *IFAC Proc. Vol.*, vol. 19, no. 9, pp. 7–14, Jun. 1986, doi: 10.1016/S1474-6670(17)57504-7.
- [12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.
- [13] C. Orhei, V. Bogdan, and C. Bonchiș, "Edge map response of dilated and reconstructed classical filters," in 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Sep. 2020, pp. 187–194. doi: 10.1109/SYNASC51798.2020.00039, WOS:000674702000028.

- [14] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851.
- [15] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision*. New York: McGraw-Hill, 1995.
- [16] G. Ramponi and A. Polesel, "A Rational Unsharp Masking Technique," *J. Electron. Imaging*, vol. 7, pp. 333–338, 1998.
- [17] K. De and V. Masilamani, "Image Sharpness Measure for Blurred Images in Frequency Domain," *Procedia Eng.*, vol. 64, pp. 149–158, 2013, doi: 10.1016/j.proeng.2013.09.086.
- [18] J. P. D. Villiers, "A comparison of image sharpness metrics and real-time sharpening methods with GPU implementations," 2010.
- [19] C. Orhei, S. Vert, M. Mocofan, and R. Vasiiu, "TMBuD: A Dataset for Urban Scene Building Detection," in *Information and Software Technologies*, Cham, 2021, pp. 251–262. doi: 10.1007/978-3-030-88304-1_20.
- [20] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features," in *Computer Vision – ECCV 2012*, Berlin, Heidelberg, 2012, pp. 214–227. doi: 10.1007/978-3-642-33783-3_16.
- [21] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, Berlin, Heidelberg, 1998, pp. 137–142. doi: 10.1007/BFb0026683.
- [22] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Jun. 2006, vol. 2, pp. 2161–2168. doi: 10.1109/CVPR.2006.264.
- [23] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Trans. Math. Softw.*, vol. 3, no. 3, pp. 209–226, Sep. 1977, doi: 10.1145/355744.355745.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778. Accessed: Jan. 18, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
- [25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
- [26] H. Shao, T. Svoboda, and L. Van Gool, "Zubud-Zurich buildings database for image based recognition." Technique Report No. 260, Swiss Federal Institute of Technology, 2003.
- [27] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 2010, pp. 3304–3311. doi: 10.1109/CVPR.2010.5540039.
- [28] D. M. Chen, S. S. Tsai, V. Ch, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," 2009.
- [29] W. Zhang and J. Koščeká, "Hierarchical building recognition," *Image Vis. Comput.*, vol. 25, no. 5, pp. 704–716, May 2007, doi: 10.1016/j.imavis.2006.05.016.
- [30] N. J. C. Groeneweg, B. de Groot, A. H. R. Halma, B. R. Quiroga, M. Tromp, and F. C. A. Groen, "A Fast Offline Building Recognition Application on a Mobile Telephone," in *Advanced Concepts for Intelligent Vision Systems*, Berlin, Heidelberg, 2006, pp. 1122–1132. doi: 10.1007/11864349_102.