UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

Decizie nr. 1092 **din** 24-07-2023

REZUMAT TEZĂ DE DOCTORAT

Ing. Andrei-Mircea RACOVIŢEANU

ÎNVĂŢAREA CU MARJĂ LARGĂ PENTRU ANALIZA IMAGINILOR

LARGE MARGIN LEARNING FOR IMAGE ANALYSIS

COMISIA DE DOCTORAT

Prof. Dr. Ing. Mihai Ciuc Univ. Politehnica din Bucureşti	Preşedinte
Prof. Dr. Ing. Corneliu Florea Univ. Politehnica din Bucureşti	Conducător de doctorat
Prof. Dr. Ing. Cătalin Căleanu Univ. Politehnica din Timişoara	Referent
Conf. Dr. Ing. Ioan Buciu Universitatea din Oradea	Referent
Prof. Dr. Ing. Constantin VERTAN Univ. Politehnica din Bucureşti	Referent

BUCUREŞTI 2023

Cuprins

Capitolul 1

Introducere

1.1 Prezentarea domeniului tezei de doctorat

În ultimii ani, domeniul inteligenţei artificiale şi a vederii computerizate au cunoscut o evoluţie foarte rapidă. Accesul mai facil la un volum mare de date împreună cu algoritmii de învăţare automată au contribuit la dezvoltarea sistemelor care imită capabilităţile umane. De progesul domeniilor amintite anterior beneficiază mai multe industrii cum ar fi: medicina, finanţele sau industria de divertisment.

Cele 2 probleme tratate în cadrul lucrării sunt recunoaşterea expresiilor faciale şi regăsirea imaginilor cu conţinut similar. Detecţia expresiilor faciale se concentrează pe recunoaşterea automată a emoţiilor oamenilor. Aceasta implică dezvoltarea unor sisteme care pot interpreta şi înţelege stările si reacţiile emoţionale ale unei persoane bazate pe mişcările faciale. Regăsirea imaginilor cu conţinut similar implică utilizarea unor tehnici de căutare şi extragere a imaginilor dintr-un volum mare de date după conţinutul vizual.

1.2 Scopul tezei de doctorat

Scopul principal al tezei este găsirea unor soluţii mai eficiente pentru recunoaşterea expresiilor faciale şi regăsirea de imagini similare. Tehnicile propuse s-au bazat pe reţele convoluţionale cât şi pe introducerea unor noi funcţii de cost şi metode de augmentare.

Reţelele convoluţionale au nevoie de cantităţi mari de date care de cele mai multe ori nu sunt uşor de achiziţionat. Pentru a combate acest impediment, se pot folosi tehnici bazate pe învăţarea semi-supervizată sau adaptarea de domeniu. Din acest motiv, soluţiile propuse depăşesc categoria algoritmilor "supervizaţi" si au fost incluse în categoria "nesupervizată".

Prima problemă abordată este cea a expresiilor faciale. În acest caz, au fost folosite atât expresii faciale discrete cţ şi mişcări faciale cunoscute ca "unităţi de acţiune". Chiar dacă mişcările faciale sunt mai uşor de înţeles, unităţile de acţiune sunt o reprezentare mai obiectivă şi mai greu de confundat. Pentru a îmbunătăţi rezultatele a fost propusă o nouă

funcţie de cost cu rol de grupare al spaţiului descriptiv furnizat de reţeaua convoluţională. În plus, au fost testate şi noi metode de augmentare si regularizare.

Cealaltă temă studiată a fost regăsirea imaginilor cu conţinut similar. În această situaţie conţinutul vizual al imaginilor poate fi foarte complex, motiv pentru care descriptorii obţinuţi cu o reţea convoluţională se pot confunda destul de uşor între ei. Aceeaşi funcţie de cost a fost testată pentru a verifica dacă o organizare mai bună a spaţiului descriptorilor îmbunătăţeşte numărul de imagini similare returnate pentru o poză de referinţă.

1.3 Conţinutul tezei de doctorat

Lucrarea a fost împărţită în 7 capitole. Capitolele cu numarul 2,3 şi 4 sunt prezentate conceptele teoretice care au reprezentat fundamentul rezultatelor experimentale. Acestea sunt aduse in prim plan în capitolele 5 şi 6, urmând ca ultimul capitol să fie rezervat concluziilor.

Capitolul 2 redă informaţii despre reţelele convoluţionale. Aici sunt prezentate principalele tipuri de straturi, funcţiile de cost, metodele de optimizare şi principalele arhitecturi folosite. În capitolul 3 sunt aduse în discuţie informaţii despre tipurile de învăţare automată. Accentul este pus pe învăţarea semi-supervizată/învăţarea prin transfer şi pe tehnicile de augmentare folosite. Capitolul 4 se concentrează pe funcţiile de cost care au rol de grupare al spaţiului descriptiv. Sunt prezentate conceptele matematice, funcţionalitatea, dar şi o scurtă comparaţie a acestora. Capitolele 5 şi 6 cuprind rezultatele obţinute în cadrul tezei pentru cele 2 teme abordate, iar capitolul 7 este destinat concluziilor.

Capitolul 2

Reţele Convoluţionale

Reţelele neuronale convoluţionale sunt o clasă de modele de învăţare profundă care au avut un succes deosebit în sarcini de vedere computerizată, cum ar fi clasificarea imaginilor, detectarea obiectelor sau segmentarea imaginilor. Acestea sunt concepute pentru a învăţa și extrage automat caracteristici relevante din datele de intrare, făcându-le potrivite pentru sarcini care implică imagini.

2.1 Straturile reţelelor convoluţionale

Ideea de bază din spatele reţelelor convoluţionale este folosirea mai multor tipuri de straturi şi procedee matematice astfel încât sa fie surprinse trăsături cât mai relevante pentru datele de intrare. Primul strat relevant este evident cel convoluţional unde este aplicată efectiv operaţia de convoluţie cu ajutorul unor filtre care conţin diferite ponderi. Aici apare şi conceptul de conectivitate locală care înseamnă că nu există conexiuni între toţi neuronii straturilor consecutive.

Următoarele straturi importante sunt cele de sub-eşantioanre. Ele au rolul de a reduce dimensiunea hărţilor de trăsături produse de operaţiile convoluţionale. În acest mod efortul de calcul este mult mai mic, iar sistemul are o putere de generalizare mai bună. Cea mai cunoscută variantă de sub-eşantionare este *max-pooling* care ia în considerare valoarea maximă dintr-o vecinătate.

Fenomenul de supraînvăţare poate apărea destul de des la reţelele convoluţionale dacă nu sunt utilizate o serie de straturi care se numesc de regularizare. Un astfel de strat este cel de *dropout* [1] care implică eliminarea la întâmplare al unui anumit procent din conexiunile neuronale dintre straturi. Acestea sunt eliminate doar în procesul de antrenare, nu şi în cel de testare. O alta variantă este normalizarea loturilor de date (*batch normalization*) [2] care constă în normalizarea activărilor fiecăriu strat. Astfel timpul de antrenare este redus, iar cazurile în care activările între straturi consecutive sunt complet diferite sunt eliminate. Ultimele straturi sunt cele dens conectate care sunt folosite adesea ca descriptori sau straturi decizionale.

2.2 Procesul de învăţare

Un sistem de învăţare automată foloseşte o serie de functii matematice care se numesc fuctii de cost pentru a măsura cât de departe se situează predicţiile acestuia de etichetele de referinţă. Există 2 tipuri de probleme în funcţie de natura etichetelor: *clasificare* (etichete discrete) sau *regresie* (etichete continue). Pentru problemele de clasificare cea mai utilizată funcţie de cost este *entropia încrucişată*, iar pentru problemele de regresie este *eroarea pătratică medie*.

Pe parcursul antrenării sistemul învaţă sa reducă erorile printr-un algoritm de optimizare matematică (*optimizator*). Acesta ajustează ponderile după fiecare iteraţie În funcţie de influenţa fiecărei ponderi la eroarea totală representata de funcţia de cost. Tot acest proces de ajustare al ponderilor care depinde de eroarea totală se numeşte propagarea înapoi a erorii (*backpropagation*).

2.3 Arhitecturi de reţele convoluţionale

Odată ce reţelele convoluţionale au devnit mai folosite s-au căutat o serie de arhitecturi standard care să poată fi folosite indiferent de natura problemei ce trebuie rezolvată. Prima arhitectură care a obţinut rezultate impresionante a fost AlexNet [3]. Odată cu aceasta a fost introdusă şi o nouă funcţie de activare care anulează ponderile negative.

O altă familie cunoscută de arhitecturi este VGG [4] care a reuşit să crească adâncimea reţelelelor prin adăgarea de noi straturi. Spre deosebire de AlexNet, filtrele de convoluţie au o dimensiune mai mică, iar la operaţiile de sub-eşantionare ferestrele nu se mai suprapun. Totuşi creşterea tot mai mare a numărului de straturi a favorizat apariţia unui alt fenomen supărător care se numeşte "dispariţia gradienţilor" (*vanisging gradients*).

Arhitecura ResNet [5] a fost propusă pentru a combate fenomenul amintit mai sus. Principala componentă inovativă a fost blocul rezidual. Spre deosebire de celelalte arhitecturi amintite în arhitecturile reziduale intrarea unui strat este copiată la intrarea unor straturi superioare. În acest fel s-a putut creşte din ce în ce mai mult adâncimea reţelelor fără a mai avea un impact negativ asupra performanţei. După ce s-a constatat că creşterea excesivă a adâncimii nu mai aduce beneficii semnificative au fost propuse si alte variante de ResNet. Wideresnet [6] propune o creştere a hărţilor de trăsături mărind astfel lăţimea straturilor convoluţionale.

Capitolul 3

Concepte de învăţare automată

3.1 Tipuri de învăţare

În general algoritmii de învăţare automată pot fi împărţiţi în 3 mari categorii după natura etichetelor datelor de intrare. Dacă datele conţin etichete algoritmii se numesc supervizaţi, altfel se numesc nesupervizaţi. Mai exista si cea de-a treia categorie a algoritmilor semi-supervizaţi care folosesc de obicei o cantitate mai mică de date adnotate şi o porţiune însemnată de date neadnotate. Aceştia din urmă au fost preponderent folosiţi la partea experimentală.

3.2 Învăţare prin transfer/Adaptare de domeniu

Datorită datelor insuficiente şi incorect etichetate, tehnicile de învăţare prin transfer si adaptare de domeniu au devenit din ce în ce mai ofertante. Invăţarea prin transfer implică folosirea unui sistem pre-antrenat pentru o problemă similară sau diferita faţă de cea pe care a fost antrenat iniţial. Cu alte cuvinte, această tehnică încearcă să crească performanţele pentru o sarcină nouă folosind caracteristici învăţate anterior.

De cealaltă parte, adaptarea de domeniu presupune găsirea unei legături între două domenii care au o distribuţie diferită a datelor. De obicei este antrenat un sistem pentru un domeniu iniţial ca mai apoi să fie adaptat la un domeniu ţintă prin diverse tehnici de ajustare cum ar fi: ponderarea funcţiilor de cost sau aliniera trăsăturilor. Rezultatul final este creşterea puterii de generalizare prin reducerea diferenţelor structurale dintre cele două domenii.

3.3 ALgoritmi semi-supervizaţi

3.3.1 Pseudo-etichetare

În momentul în care un sistem antrenat pe date care conţin etichete este aplicat pe o serie de eşantioane care nu deţin această informaţie se obţine pseudo-etichetarea [7]. Practic, se obţin etichetele pentru setul de date neadnotat folosind un sistem pre-antrenat. Apoi, se adaugă datele cu etichetele generate la cele iniţiale si se reantrenează tot ansamblul.

3.3.2 Mean-Teacher

Algoritmul *Mean-Teacher* [8] este un algoritm semi-supervizat care foloseşte 2 reţele similare. Prima este considerată "student" şi are rol de clasificare. Cea de-a doua este reţeaua "profesor" care trebuie să reproducă cât mai bine ieşirea reţelei "student". Pe parcursul procesului de antrenare, unul dintre obiectivele principale este reducerea diferențelor de distribuție între reţeaua "student" și reţeaua "profesor".

3.3.3 MixMatch

MixMatch [9] este un algoritm care combină mai multe paradigme semi-supervizate pentru a maximiza performanţa. Acesta foloseşte metoda *MixUp* [10] prin care se generează noi eşantioane şi noi etichete. Se utilizează o funcţie de cost care ia în calcul şi componenta de informaţie furnizată de noile date generate.

3.4 Metode de augmentare

Ţinând cont că reţelele convoluţionale au nevoie de cantităţi mari de date s-au căutat variante care să rezolve această problemă. Metodele de augmentare se folosesc pentru a mări artificial numărul de eşantioane folosite la antrenare.

Cele mai uzuale metode de augmentare sunt legate de distribuţia spaţială a elementelor dintr-o imagine. Astfel se pot genera noi eşantioane prin operaţii de rotire, oglindire sau redimensionare. Se poate acţiona şi la nivel de pixel prin tehnici de filtrare sau contrastare.

Tehnica de augmentare *MixUp* [10] ajută la generarea de date noi printr-o combinaţie liniară a unei perechi aleatoare de eşantioane existente în baza de date folosită la antrenare. Pentru aceste date se creează etichete folosind aceeaşi metodă. Totuşi, fiind dificil de interpretat ca valoare unica combinaţia liniară a 2 etichete discrete s-a folosit distribuţia probabilistică obţinută.

Capitolul 4

Metode de structurare al spaţiului descriptiv

Pentru probleme complexe este posibil ca spaţiul descriptiv furnizat de reţelele convoluţionale să conţină exemple care se suprapun destul de mult. Din acest motiv, s-au realizat cercetări în direcţia unor funcţii de cost noi care să organizeze mai eficient caracteristicile obţinute.

4.1 Funcţia de cost centrică

Una din cele mai cunoscute astefel de funţii este cea centrică [11]. Principiul după care se ghideaza este minimzarea distanţelor între punctele care sunt încadrate în aceeaşi categorie. Ecuaţia 4.1 descrie matematic funţia de cost centrică, unde e_i este descriptorul obţinut de pe stratul dens conectat înainte de cel decizional, iar c_i este centroidul asociat eşantionului curent.

$$L_C = \frac{1}{2}\sum_{i=1}^{N} \|e_i - c_{l_i}\|_2 \tag{4.1}$$

Pentru a obţine un spaţiu mai aerisit, poziţia centroizilor este recalculată după fiecare iteraţie conform 4.2. Δc_k^i este media datelor care aparţinde clasa i în lotul de date curent, iar α este un parametru subunitar care temprează o posibilă influenţă negativă la noua poziţie a centroizilor a unor eşantionae etichetate eronat. Deoarece nu are rol de decizie/clasificare, funcţia centrică se poate folosi doar cu o funcţie de cost de acest tip.

$$c_k^{i+1} = c_k^i - \alpha \Delta c_k^i \tag{4.2}$$

4.2 Funcţia de cost insulară

Spre deosebire de funcţia de cost amintită în secţiunea precedentă, funcţia insulară [12] aduce un nou plus. Pe lângă minimizarea varianţei în aceeaşi clasă, aceasta încearcă să maximizeze distanţa între centroizii asociaţi fiecărei etichete. Ecuaţia 4.3 descrie formula matematică, unde L_C este funcţia de cost centrică, iar termenul al doilea cumulează distanţele unghiulare între centroidul curent şi toate celelalte clase. La fel ca în cazul precedent se foloseşte cu o funcţie de cost de clasificare.

$$L_{IL} = L_C + \lambda_1 \sum_{c_i \in C} \sum_{c_j \in C, c_i \neq c_j} \left(\frac{c_i \cdot c_j}{\|c_i\|_2 \|c_j\|_2} + 1 \right) \tag{4.3}$$

4.3 Funcţia de cost circulară

Funcţia de cost circulară [13] implică normalizarea eficientă a descriptorilor astfel încât aceştia să să poată fi interpretaţi din punct de vedere geometric ca un cerc. În ecuaţia 4.4 F_{x_i} este descriptoul asociat penultimului strat dens conectat, iar R este norma care este vizată pentru fiecare eşantion. Din punct de vedere geometric este asocită cu raza unui cerc.

$$L_R = \frac{\lambda}{2m} \sum_{i=1}^{m} (\|F_{x_i}\|_2 - R)^2 \tag{4.4}$$

4.4 Funcţia de cost marjă largă

Funcţia de cost centrică ia în calcul doar distanţa în interiorul claselor, nu şi distanţa între punctele care se situează în clase distincte. Funcţia insulară combate acest neajuns, dar bazandu-se pe distanţa unghiulară are o serie de limitări. Dacă unghiul dintre 2 grupuri de eşantioane este foarte mic acestea nu se pot separa suficient. Funcţia circulară impune o reprrezentare geometrică a descriptorilor sub forma unui cerc maximizând astfel distanţele unghiulare. Totuşi, dacă 2 clase se suprapun în reprezentarea iniţială există şanse mari ca acestea să coicidă şi în reprezentarea circulară.

Luând în considerare limitările amintite în paragraful anterior, a fost propusă o nouă funcţie de cost care să satisfacă nevoile actuale. Aceasta a fost intitulată funcţia de marjă largă [14] şi impune folosirea unei distanţe euclidiene între fiecare eşantion şi centroizii asociaţi celorlalte clase. În acest mod sunt evitate limitările impuse de de funcţiile de cost anterior prezentate.

Ecuaţia 4.5 redă definirea matematică a acesti funcţii. e_i este descriptorul reprezentat de penultimul strat dens conectat, c_j reprezintă centroidul asociat descriptorului e_i, iar c_k reprezintă toţi centroizii diferiţi de centroidul lui e_i.

$$\mathbf{L_{LM}} = \sum_{i=1}^{N} \left(\left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{k=1, k\neq j}^{C} \left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}_k}{\|\mathbf{c}_k\|_2} \right\|_2 \right) \tag{4.5}$$

În Figura 4.1 este prezentată pe scurt funcţionalitatea funcţiei de marjă largă. Având la dispoziţie 3 eşantioane (X_A;X_B;X_C) cu etichetele A,B si C, acestea vor fi aduse mai apropae de centroizii claselor de care aparţin (minimizarea distanţelor marcate cu sageţi continue) şi vor fi depărtate de centroizii celorlate clase (maximizarea distanţelor marcate cu linie punctată).

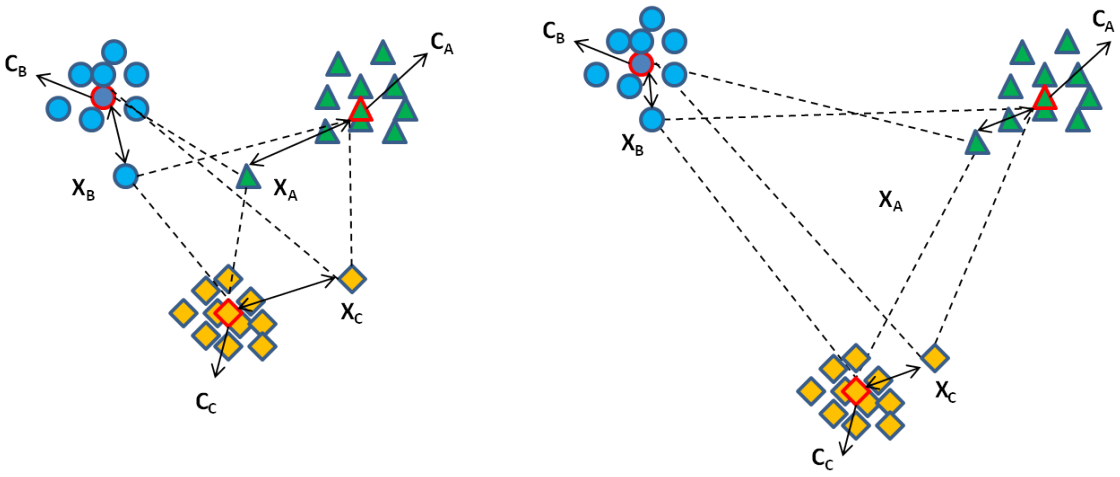

Figura 4.1 Funcţionalitatea costului de marjă largă. Stânga: Înainte de marja largă. Dreapta: După marja largă

Capitolul 5

Analiza expressilor faciale

Acest capitol este cel mai cuprinzător din întreaga teză şi reprezintă efortul mai multor ani de cercetare. Printre sarcinile abordate se regăsesc recunoaşterea de expresii discrete, dar şi detecţia de mişcări faciale. Pentru îmbunătăţirea performanţei s-au utilizat funcţii noi de cost şi metode inovative de augmentare şi regularizare.

5.1 Cuantificare expresiilor faciale

Recunoaşterea expresiilor faciale nu este o sarcină uşoară nici pentru oameni. Aceasta este cu atât mai dificilă pentru un sistem de învăţare automată. De-a lungul timpului s-au căutat mai multe modele care să definească mai convenabil această problemă.

Primul model foarte folosit în acest domeniu este cel al expresiilor discrete definit de Ekman [15]. Acesta are la bază 6 expresii fundamentale : frică, dezgust, fericire, furie, surprindere, tristeţe. De obicei, la acestea se mai adaugă şi expresia neutră. Deşi această variantă este simplu de folosit, s-au căutat şi alte modele care să definească emoţiile într-o manieră mai obiectivă. Astfel, a fost introdus *sistemul de codare al unităţilor de acţiune* care presupune că fiecare exprsie facială este formată dintr-o serie de activări ale unor muşchi faciali specifici (unităţile de acţiune).

Sistemul cuprinde un număr de 43 de unităţi de acţiune care au fost divizate în funcţie de poziţia la nivelul feţei. Există mişcări faciale asociate celor mai importante elemente faciale cum ar fi: ochii. gura sau sprâncenele. Deşi este un model care lasă mai puţin loc de interpretări, este necesar un personal destul de calificat care să recunoască cele mai fine mişcări faciale. Mai există şi modele mai complexe care iau în calcul intensitatea expresiei sau dacă este pozitivă sau negativă. Un asemenea model a fost propus de Russel [16] şi ia în considerare emoţiile compuse.

5.2 Provocări în recunoaşterea expresiilor faciale

După cum s-a demonstrat şi în secţiunea 5.1 problema expresiilor faciale este una foarte dificil de rezolvat. Mai mult, există şi o multitudine de provocări care pot afecta performanţa de recunoaştere. Un prim motiv este volumul de date inconsistent. Chiar dacă poate că este facil să obţii multe imagini cu expresii faciale, procesul de etichetare poate fi lung şi costisitor. Nu este o misiune uşoară să fie găsiţi adnotatori umani profesionişti. Din această cauză există încă multe imagini cu etichete greşite.

Un alt motiv este timpul scurt de apariţie al expresiilor faciale. De cele mai multe ori mişcaările faciale care alcătuiesc emoţiile sunt foarte fine chiar şi pentru computer. De asemenea, există mai multe expresii care sunt uşor de confundat între ele cum ar fi frica şi surprinderea. Ambele emoţii implică deschiderea gurii şi ridicarea sprâncenelor.

5.3 Legătura cu literatura de specialitate

Fiind o sarcină greu de rezolvat, recunoaşterea expresiilor faciale a atras în ultima perioadă un număr mare de cercetători, motiv pentru care articolele pe această temă au crescut considerabil. În lucrări precum [17, 18] a fost abordată recunoaşterea expresiilor fundamentale (secţiunea 5.1) cu reţele convoluţionale .

Du [19] şi Zhang [20] au observat inconsistenţa datelor de antrenare şi au migrat spre soluţii din zona învăţării semi-supervizate. A fost abordată şi recunoaşterea unităţilor de actiune. Corneanu [21], Zhao [22] sau Benitez [23] împreună cu colaboratorii sunt doar o parte din cei care au detectat mişcările faciale şi intensitatea acestora. Adaptarea de domeniu şi învăţarea prin transfer au fost abordate în [24–26].

5.4 Soluţii de detecţie facială

Cele mai multe seturi de date conţin imagini cu feţe sau expresii faciale care nu sunt încadrate convenabil. Informaţia legată de fundal care se regăseşte în majoritatea imaginilor este redundantă şi nu este folositoare reţelelor convoluţionale. În plus, prin decuparea feţelor din pozele iniţiale se micşorează dimensionalitatea care contribuie la un timp mai redus de antrenare.

Din multitudinea de soluţii existente se remarcă algoritmii Viola-Jones [27] şi MT-CNN [28]. Metoda lui Viola-Jones foloseşte o serie de trăsături la diferite scale pentru a surprinde feţele din imagini oricare ar fi dimensiunea acestora. Se mai foloseşte şi imaginea integrală pentru accelerarea procesului de detecţie care a permis identificarea facială în timp real.

De cealalta parte MTCNN [28] este o tehnică bazată pe reţelele convoluţionale. Aceasta poate fi organizat în 3 etape. Se detectează mai întâi feţele, se potrivesc ferestrele

de încadrare şi la final se recunosc elementele faciale necesare pentru o aliniere cât mai corectă a feţei.

5.5 Baze de date

Pentru a face faţă cerinţelor impuse de metodele semi-supervizate testate în partea experimentală au fost folosite mai multe seturi de date diverse. S-a abordat problema recunoaşterii de expresii fundamentale,dar şi a detecţiei de mişcări faciale (unităţi de acţiune). Mai multe informaţii despre fiecare în parte se regăsesc în secţiunile ce urmează.

5.5.1 Baze de date pentru expresii faciale

FER/FER+. FER2013 [29] şi FER+ [30] reprezintă 2 seturi de date care oferă imagini cu expresii faciale fundamentale. FER+ este o extindere a FER2013 în care s-au corectat o serie de etichete care au fost atribuie incorect în prima versiune. Dispune de un număr de aproximativ 35000 de imagini achiziţionate în condiţii naturale.

Megaface. Megaface [31] este o bază de date mult mai cuprinzătoare care conţine aproximativ 1 milion de imagini achiziţionate în condiţii naturale. Nu dispune de etichete, motiv pentru care a fost folosită ca porţiune neetichetată de date pentru experimentele semi-supervizate.

RAF-DB. RAF-DB [32] seamană cu FER deoarece conţine expresii discrete. Cu toate acestea, dispune de un număr mai mic de fotografii obţinute în condiţii de laborator. Adnotarea imaginilor a fost realizată de un număr de 40 de persoane.

Facial expression in children. În cadrul acestei lucrări a fsot studiată şi recunoaşterea expresiilor faciale la copii. CAFE [33] este unul din cele mai cunoscute seturi de date pe această temă. LRIS [34] este un alt set care compensează limitările CAFE prin creşterea numărului de imagini şi diversificarea emoţiilor.

5.5.2 Baze de date pentru detecţia unităţilor de acţiune

CK+. CK+ [35] este un set de date care conţine atât etichete discrete cât şi unităţi de acţiune. Este prezentat în această secţiune pentru că a fost folosit doar pentru recunoaşterea de unităţi de acţiune. Imaginile sunt organizate pe subiecţi şi fiecare secvenţă conţine cadre începând de la expresia neutră până la intensitatea maximă a expresiei. Ultimele poze din fiecare secvenţă sunt adnotate şi la nivel de unităţi de acţiune.

Emotionet. Emotionet [23] este un set de date care conţine aproximativ 1 milion de imagini. Spre deosebire de Megaface [31], dispune de 50000 de imagini etichetate la nivel de unităţi de acţiune. Imaginile sunt achiziţionate în condiţii naturale. Etichetele pentru mişcările faciale sunt binare.

DISFA. În comparaţie cu Emotionet [23], DISFA [36] oferă adnotări şi pentru intensitatea unităţii de acţiune. Astfel, etichetele sunt cuprinse între 0 (unitatea de acţiune nu este activă) si 5 (unitatea de acţiune are intensitatea maximă). Conţine 130000 de imagini captate în condiţii de laborator împărţite în 27 de subiecţi.

5.6 Recunoaşterea expresiilor faciale

Acest capitol s-a ocupat de îmbunătăţirea rezultatelor pentru detecţia expresiilor faciale folsind mai multe instrumente noi sau deja cunoscute în literatură. Printre acestea se număra învăţarea semi-supervizată/învăţarea prin transfer, funcţiile de cost cu rol de structurare al spaţiului descriptorilor (centrică - Secţiunea 4.1, circulară - Secţiunea 4.3, marjă largă - Secţiunea 4.4), dar şi metode de augmentare precum MixUp (Secţiunea 3.4).

5.6.1 Recunoaşterea expresiilor faciale cu funcţie de cost de marjă largă şi pseudo-expresii

Având ca motivaţie succesul funcţiilor de cost centrice şi insulare, în această secţiune a fost explorat potenţialul funcţiei de marjă largaă care a fost propusă (Secţiunea 4.4).

Pentru acest scenariu a fost folsită o arhitectură Alexnet (Secţiunea 2.3) care a fost antrenată pentru detecţia simultană a expresiilor faciale discrete şi a unităţilor de acţiune. La nivel de arhitectură au fost folosite 2 straturi de decizie conectate în paralel din penultimul strat dens conectat.

S-a încercat folosirea la antrenare a mai multor seturi de date care nu conţin acelaşi tip de etichetă. De aceeaa a fost necesară o adaptare de domeniu plecând de la unităţi de acţiune la expresii faciale discrete. Legătura dintre cele 2 categorii s-a realizat cu un set de ecuaţii care descriu expresiile fundamentale ca fiind o sumă de unităţi de acţiune active simultan. Rezultatul a constat în obţinerea unor pseudo-expresii.

A fost nevoie de aceste pseudo-expresii, deoarece funcţia de cost de marjă largă are nevoie de conceptul de centroizi (clase discrete). Detecţia de unităţi de acţiune este o problemă multi-clasă deoarece pot fi active mai multe mişcări faciale în acelaşi timp. Chiar dacă etichetele corespunzătoare unităţilor de acţiune sunt discrete, faptul amintit mai devreme poate conduce la un număr excesiv de clase cu reprezentare redusă.

Rezultatele pentru recunoaşterea expresiilor faciale pot fi urmărite în tabelul 5.1, în timp ce performanţa pentru unităţile de acţiune este redată în tabelul 5.2. Se poate observa că funcţia de cost de marjă largă obţine rezultate mai bune decât funcţia centrică şi insulară.

Metodă	Învăţare	Med. Ac.	Ac.
AlexNet - [32]	Superv	55.60	68.90
AlexNet + Feat.Sel.Net [17]	Superv	72.46	81.10
AlexNet + Island loss [12]	Superv	57.1	75.08
AlexNet + Center loss [11]	SSL	63.15	78.81
AlexNet + Island loss [12]	SSL	64.53	78.81
AlexNet + LM loss [14]	SSL	67.26	79.85

Tabel 5.1 Acurateţea obţinută pentru setul de date RAF-DB. Tabel din [14]

Metodă	Înv	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU17	AU20	AU25	AU26	AU43	Med. redus	Med. total
AlexNet [3]	Sv	24.2	n/a	34.7	**39.5**	73.1	n/a	86.8	n/a	n/a	88.5	45.6	n/a	56.1	n/a
AlexNet cen. loss [11]	Sv	**34.4**	30.3	55.3	33.3	69.10	**46.1**	79.3	27.8	**32.3**	84.4	43.2	48.8	57.9	48.8
AlexNet +WSC [22]	SSL	25.3	n/a	34.5	39.3	**75.6**	n/a	**87.4**	n/a	n/a	**88.8**	47.4	n/a	57.0	n/a
AlexNet + Isl. loss [12]	T.	30.4	29.5	**56.7**	30.6	66.7	44.1	77.3	26.7	23.8	83.9	47.3	43.9	56.14	46.7
AlexNet + LM loss [14]	T.	34.1	**31.1**	56.6	33.9	71.0	45.1	78.1	**30.9**	25.3	83.8	**50.9**	47.2	58.33	49.0

Tabel 5.2 Scorul F1 (%) pentru detecţia de unităţi de acţiune pentru baza de date Emotionet. Învăţarea este fie supervizată, semi-supervizată sau prin transfer. "Med. redusa" este media pentru setul redus de AUs:1,4,5,6,12,25,26, Med. total este media pentru setul întreg. Tabel din [14]

5.6.2 Margin-Mix

Algoritmul *Margin-Mix* [37] combină funcţia de cost de marjă largă cu o serie de tehnici de augmentare cum ar fi MixUp [10]. Aşa cum a fost amintit şi în secţiunea 3.4 tehnica MixUp este aplicată pentru învăţarea pur supervuzată. Deşi combinaţia liniară între 2 imagini din setul de date este posibilă, rezultatul nu va putea fi atribuit unei clase în lipsa unor etichete iniţiale.

Aici intervine conceptul de marjă largă care este folosit pentru etichetarea noilor exemple formate cu MixUp folosind descriptorii formaţi de o reţea convoluţională. Pentru a minimiza efectul suprapunerii datelor în spaţiul descriptiv atribuirea etichetelor s-a realizat folosind o tehnică Fuzzy [38]. În acest fel unui eşantion i-a fost atribuită o distribuţie de probabilitate pentru fiecare clasă, nu doar o etichetă unică. Mai departe, aceste noi eşantionane împreună cu etichetele prezise vor putea fi folosite mai departe pe parcursul antrenării.

De-a lungul experimentelor a fost folosită o arhitecturıa Wide-ResNet [6]. Performanţa obţinută pentru seturile de date standard: CIFAR10, STL-10, SVHN se regăseşte în Tabelul 5.3. Rezultatele pentru RAF-DB [32] au fost organizate în Tabelul 5.4. În

primul tabel rezultatele sunt asemănătoare cu cele din literatura de specialitate. Mai sugestive sunt rezultatele din cel de-al doilea tabel. Se remarcă o performanţă considerabil mai bună pentru *Margin-Mix* în comparaţie cu alte metode pur supervizate. Este important de semnalat că pe măsură ce cantitatea de date etichetate este tot mai mică metoda propusă are cele mai bune rezultate. Totuşi, atunci când este folosit tot setul de date (ultima coloană din tabelul 5.4) numerele sunt similare, ceea ce demonstrează intr-adevăr că Margin-Mix este o soluţie de luat în calcul atunci când setul de date nu conţine suficiente adnotări.

	SVHN		**STL**	
Metodă/Etichete	1000	4000	1000	5000
Supervizat [6]	–	12.84	–	–
Model Π [39]	8.06	5.57	17.41	39.19
VAT [40]	5.63	18.68	11.05	–
MeanTeacher [8]	5.65	3.39	10.36	–
ICT [41, 6]	3.53	–	7.66	–
MixMatch [9]	3.27	2.89	10.18	5.59
MarginMix [37]	3.35	8.33	**9.85**	5.80

Tabel 5.3 Eroarea obţinută pentru seturile de date STL şi SVHN cu arhitectura Wide - ResNet.[37]

Metodă/Etichete	320	400	1000	4000	
Supervizat	nc	26.75	35.25	55.66	85.58
Supervizat [32]	–	–	–	–	84.13
MeanTeacher [8]	nc	28.83	36.53	60.36	–
MixMatch [9]	35.60	42.25	60.37	65.24	–
MarginMix [37]	**40.55**	**45.75**	**66.47**	**70.68**	85.36

Tabel 5.4 Acurateţea obţinută pentru RAF-DB cu arhitectura Wide - ResNet. Linia de sus expune numărul de exemple etichetate luate în considerare. nc - nu s-a atins convergenţa [37]

5.6.3 Injecţia aleatoare în gradient pentru un transfer eficient în recunoaşterea expresiilor faciale

Tehnica propusă în [42, 43] are în componenţă învăţarea semi-supervizată şi tehnica de etichetale amintită în secţiunea 3.3.1 (Pseudo-etichetarea). Pseudo-etichetarea este simplu de utilizat, ea constând în folosirea unui sistem preantrenat pe o problemă supervizată pentru a eticheta o serie de eşantioane fară etichete.

Cu toate acestea, Pseudo-etichetarea pleacă de la premisa că datele etichetate, cât şi cele neetichetate au o distribuţie similară. Această idee nu este adevarată de cele mai

multe ori şi poate avea cauze nefaste asupra performanţei. În Figura 5.1 este expusă schematic modalitatea de funcţionare a metodei discutate. Pentru cazul supervizat (a) lucrurile sunt clare, graniţa de separare fiind raportată la datele existente. Atunci când apar şi date neadnotate (cercurile albastre) Pseudo-etichetarea obligă sistemul să devină prea sigur pe etichetele furnizate, chiar dacă ele nu sunt corecte. În acest caz, graniţa poate arăta ca la un sistem predispus supraînvăţării. Dacă se foloseşte injecţia aleatoare de gradient (IAG) se induce indirect un grad aleator de incertitudine la predicţie, motiv pentru care graniţa riscă mai puţin să fie focusată prea mult asupra punctelor incerte.

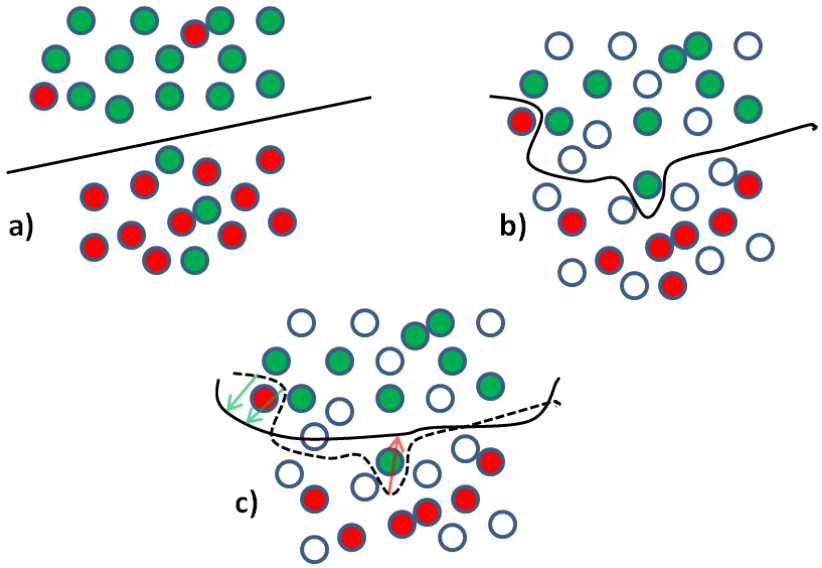

Figura 5.1 Graniţa de separaţie pentru a) Separare ideală supervizată b) Pseudo-etichetare - semi-supervizată c) Injecţie aleatoare a gradientului (IAG) - Învăţare prin transfer

Funcţia matematică prin care s-au alterat gradienţii este definită ca:

$$g(n,\lambda) = \begin{cases} \frac{\lambda n}{50}, & n < 50 \\ 0, & n \geq 50 \end{cases} \tag{5.1}$$

unde g este o variabilă aleatoare $g : \{1, N_{epochs}\} \times [-1,1] \rightarrow [-1,1]$, λ este o valoare aleatoare subunitară din intervalul [-1;1], iar n este numărul de epoci. Această cantitate este adunată la costul curent, iar ponderile se modifică doar dacă câştigul de performanţă este suficent de mare. În această manieră s-a combătut potenţialul negativ care poate fi generat de distribuţiile diferite între seturile de date.

Rezultatele pentru baza de date RAF-DB sunt expuse în tabelul 5.5. Se observă că numerele obţinute cu injecţia aleatore în gradient sunt preponderent mai mari decât metodele alese spre comparaţie. Merită semnalat că performanţa este cu aproximatic 2-3% mai bună decât pseudo-etichetarea, ceea ce arată un transfer mai eficient de informaţie.

S-au efectuat o serie de teste şi pe baza de date LRIS [34] care conţine imagini cu expresii faciale la copii. Performanţa este redată în tabelul 5.6. Şi in această situaţie rezultatele au fost mult mai bune raportat la varianta supervizată aleasă ca reper.

	Metodă / Metrică	Med. Ac.	Ac.
SUPERVIZAT	AlexNet [32]	55.60	68.90
	VGG-16 [32]	58.22	70.53
	DLP-CNN [32]	74.20	84.13
	ResNet-18 [44]	–	80.00
	RST [45]	72.46	81.10
	gCNN [46] - VGG16	–	85.07
	ensCNN [47]	75.14	86.31
TRANSFER	AlexNet + PE	69.5	78.5
	AlexNet + TE [42]	72.3	81.50
	AlexNet + IAG [43]	72.5	82.1
	VGG–16 + PE	74.6	83.25
	VGG–16 + TE [42]	76.50	84.5
	VGG–16 + IAG - [43]	76.82	85.15
	ResNet-50 + PE	77.12	84.8
	ResNet-50 + IAG - [43]	**78.22**	**86.67**

Tabel 5.5 erformanţa pe setul RAF-DB. FSN - feature selection network, FSM - frame-to-sequence method, PE - Pseudo-etichetare, TE - Transfer etichete, RST - Reţea selecţie trăsături, Med.Ac. - AcuraTEţe medie, Ac. - Acurateţe. Metoda propusă este notată cu IAG. Cu bold sunt marcate rezultatele cele mai bune. Tabel din [42]

Metodă	Acurateţe
VGG-16 [34] - supervizat	67.2
VGG-16 - +IAG [43]	68.5
ResNet-50 - supervizat	72.3
ResNet-50 + IAG [43]	76.6

Tabel 5.6 Performanţa pentru setul de date LRIS care conţine expresii faciale la copii. Tabel din [43]

5.6.4 Detecţia unităţilor de acţiune cu funcţie de cost de marjă largă

Informaţiile din secţiunea curentă folosesc principiile de marjă largă, pseudo-expresii şi adaptare de domeniu exact ca în 5.6.1. Totuşi, experimentele şi scenariile testate au fost extinse. Aici, eforturile s-au concentrat doar pe detecţia unităţilor de acţiune [48].

Arhitecturile folosite în acest caz au fost preponderent cele din familia ResNet [5]. La fel ca în secţiunea 5.6.1 s-a realizat detecţia de unităţi de acţiune folosind şi informaţia furnizată de stratul decizional asociat pseudo-expresiilor. Astfel au fost

folosite 3 funcţii de cost diferite asociate pentru fiecare problemă în parte : entropia încrucişată binară pentru predicţia mişcărilor faciale, entropia încrucişată pentru predicţia pseudo-expresiilor şi funţia de marjă largă pentru gruparea descriptorilor. Funcţia de cost totală este redată în ecuaţia 5.2. Constantele λ_1, λ_2, λ_3 sunt folosite pentru echilibrul numeric dintre cei 3 termeni ai costului final.

$$L_T = \lambda_1 L_{BCE} + \lambda_2 L_{SE} + \lambda_3 L_{LM} \tag{5.2}$$

Una dintre problemele apărute în bazele de date folosite la experimente a fost faptul că anumite unităţi de acţiune au o frecvenţă de apariţie mult mai redusă în comparaţie cu altele. Acest aspect contribuie la o detecţie mult mai redusă a acestora, deoarece la nivel descriptiv unităţile de acţiune care apar mai puţin se vor suprapune cu celelalte aşa cum se poate remarca în Figura 5.2. În figură sunt reprezentate în partea staângă pseudo-expresiile reprezentate de unităţile de acţiune care apar mai rar, iar în partea dreaptă pseudo-expresiile reprezentate de toate unităţile de acţiune existente.

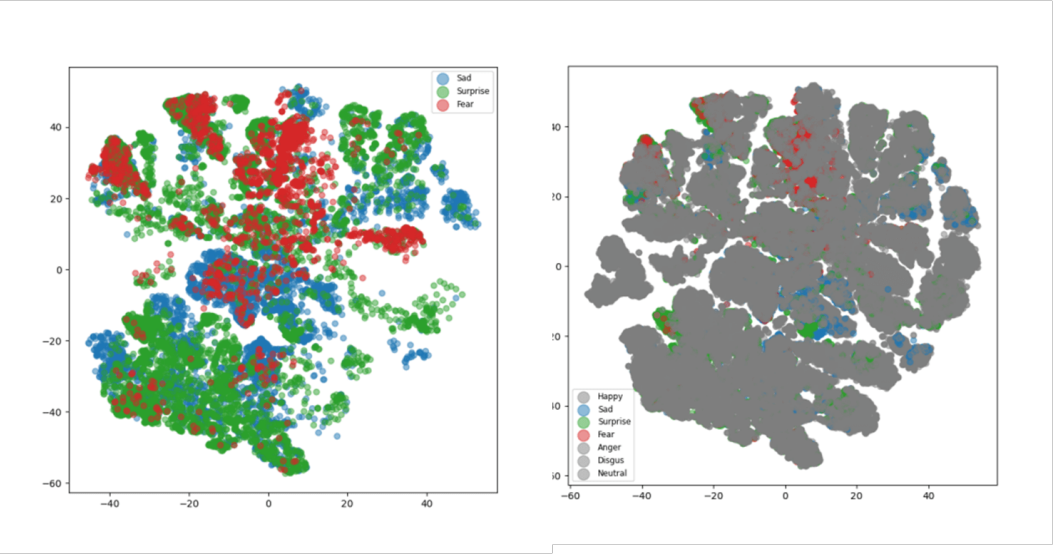

Figura 5.2 Reprezentarea la nivel descriptiv a pseudo=expresiilor. Stânga – Descriptorii asociaţi pseudo-expresiilor reprezentate de unităţile de acţiune care apar mai rar. Dreapta - Descriptorii asociaţi pseudo-expresiilor reprezentate de toate unităţile de acţiune existente. Figurua din [48]

În tabelele 5.7 şi 5.8 pot fi urmărite rezultatele obţinute pe setul de date DISFA [36] şi Emotionet [23] în comparaţie cu alte tehnici similare din literatura de specialitate. Pentru setul DISFA se poate observa că rezultatele obţinute cu funcţia de cost de marjă largă pentru unităţile de acţiune cu frecvenţă mai mică de apariţie surclasează toate celelalte tehnici din literatura de specialitate chiar dacă media totală nu este neaparat cea mai bunaă. Pe setul de date Emotionet, diferenţele se păstrează.

Un posibil motiv pentru care performanţa este mai crescută în cazul unităţilor de acţiune care apar mai puţin se datorează conceptului de marjă largă care reuşeşte să

separe mai bine la nivel descriptiv pseudo-expresiile care sunt reprezentate de unităţile de acţiune mai rare. Acest lucru se poate vedea în Figura 5.3 unde expressile formate din mişcări faciale care apar mai rar (punctele colorate) sunt mult mai compacte şi mai puţin suprapuse cu celelalte (punctele gri - unităţi de acţiune care apar mai des) atunci cand este utilizat conceptul de marjă largă.

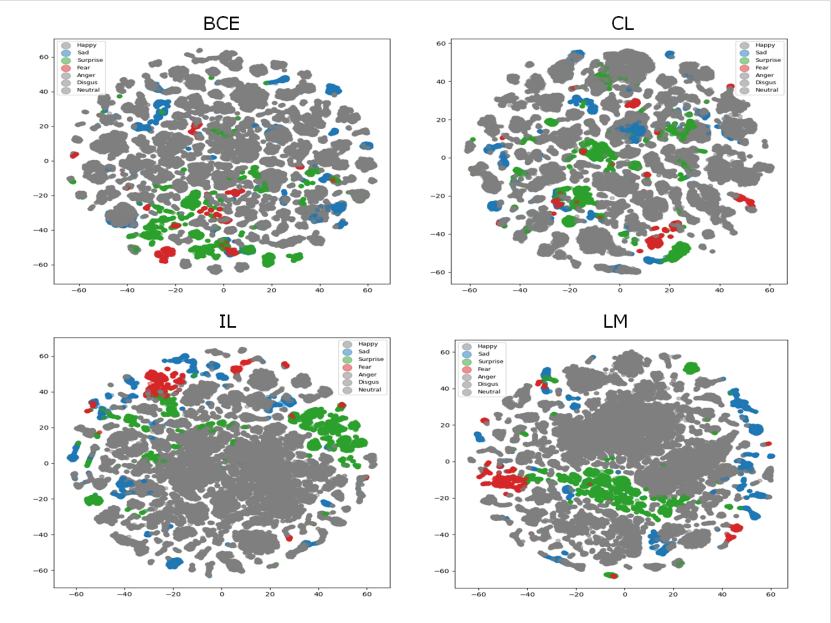

Figura 5.3 Evoluţia spaţiului descriptiv la epoca 40 pentru toate psedudo-expresiile. BCE (Entropie Binară Încrucişată) , CL(Funcţie centrică, IL (Funcţie insulară) , LM (Marjă largă) Figură din [48]

Tabel 5.7 Comparaţie cu literatura de specialitate pentru DISFA. AD reprezintă adaptare de comeniu. ML reprezintă marjă largă. F1-tot este media tuturor unităţilor de acţiune. F1-rare este media unităţilor de acţiune care apar mai rar .Cele mai bune rezultate sunt marcate cu bold. Rezultatele obţinute cu ML sunt colorate cu gri. Tabel din [48]

Metodă	F1-tot	F1-rare
AD (DISFA neetichetat) PreActRes18 - - preantrenat Funcţie de marjă largă	55.4	**51.2**
AD (DISFA neetichetat) PreActRes18 - - fară preantrenare Funcţie de marjă largă	50.4	44.7
AD (DISFA neetichetat) PreActRes18 Funţie centrică [11]	46.6	38.9
AD (DISFA neetichetat) PreActRes18 Funţie insulară [12]	47.3	40.0
AD (DISFA unlabeled) PreActRes18 Funţie circulară [13]	46.0	39.4
DRML [49]	26.7	14.2
ROI [50]	48.4	23
DSIN [21]	53.6	42.6
JAA [51]	56.0	48.2
SRERL [52]	55.9	45.6
MLT-RM [53]	**60.1**	46.6
UGN-B [54]	60.0	49.65

Tabel 5.8 Rezultate pentru Emotionet. Comparaţie cu literatura de specialitatet. F1-tot este media tuturor unităţilor de acţiune. F1-rare este media unităţilor de acţiune care apar mai rar. PsExpr = Pseudo-expresii SV = Supervizat ; AD = Adaptare Domeniu; EBI = Entropie Binară încrucişată; ML = Marjă largă; FC = Funcţie centrică; FI = Funcţie insulară; FCR = Funcţie circulară. Rezultatele cele mai bune sunt marcate cu bold. Cele mai bune rezultate obţinute cu ML sunt colorate cu gri. Tabel din [48]

Metodă	**F1-tot**	**F1-rare**
SV-BCE +PrActRes18	47.08	35.19
SV-BCE +PrActRes18 - CE(PSExpr)	48.48	36.75
SV-NM	50.16	38.90
AD - ML -Alexnet	49.04	35.96
AD - ML +PrActRes18	52.12	40.74
AD - ML +PrActRes18 Imagenet preantrenat	54.31	43.25
AD - ML+PrActRes18 DISFA preantrenat	**55.89**	**45.58**
AD - FC[11] +PrActRes18	48.92	36.83
AD - FCR[13] +PrActRes18	49.14	36.34
AD - FI[12] +PrActRes18	50.38	38.20

Capitolul 6

Regăsirea de imagini similare

Regăsirea de imagini similare este o problemă tot mai întâlnită în diverse domenii cum ar fi: motoarele de căutare de imagini sau aplicaţii medicale. De-a lungul timpului au fost folosiţi o multituine de descriptori care au avut rolul să descrie informaţia vizuală din imagini cat mai eficient. În acest capitol s-a studiat posibilitatea folosirii funcţiei de cost de marjă largă pentru a obţine o serie de descriptori mai buni folosind reţele convoluţionale.

6.1 Baza de date

Baza de date utilizată în experimente se numeşte Places365 [55] şi este alcătuită din aproximativ 1.8 milioane de imagini distribuite în 365 de clase. Imaginile conţin scene diverse. Setul de antrenare dispune de un număr situat între 3000 şi 5000 de imagini. Setul de testare are 900 de poze pentru fiecare clasă.

6.2 Legătura cu literatura de specialitate

Înainte ca reţelele convoluţionale să devină populare, pentru problema regăsirii de imagini similare s-au căutat variante foarte diverse de descriptori. Printre cei apăruţi înainte de perioada cu învăţare adâncă se regăsesc modelele binare locale [56], histograma gradientilor orientaţi [57] sau histograme de culoare. Apoi, s-a trecut la descriptori care detectează puncte similare într-o imagine cum ar fi: SIFT [58] sau SURF [59].

În ultimii ani au prins contur şi descriptori care evidenţiază mai bine informaţia vizuală cum ar fi histograma de cuvinte vizuale [60]. Evident că odată cu dezvoltarea reţelelor convoluţionale tot mai multă lume a dorit să folosească straturile dens conectate pe post de descriptori [61, 62].

Scenariu/Metrică [%]	Prec.med.-5-interogări	Prec.med.-8-interogări	Er. Top-1	Er. Top-5	Acc.	ASG PR
EI-pre-antrenat-reper	29.93	28.35	66.74	37.53	53.10	9.17
ML-fine-tunat	29.89	28.56	66.99	37.95	54.69	9.23
ML-antrenare de la zero	**31.35**	**29.98**	66.18	37.79	53.51	11.18
Resnet-152 [19]	-	-	-	-	54.74	-

Tabel 6.1 Rezultate experimentale Places 365 (Prec.med.- precizie medie, Er. - Eroare, EI- entropie încrucişată , Acc. - acurateţe, ASG PR - aria de sub graficul precizie-reamintire, ML- marjă largă).

Tabel din [63]

6.3 Metoda propusă

Pentru probelma de regăsire de imagini similare a fost folosită o reţea convoluţională ResNet care a fost antrenată pe setul de date Places 365 [55]. După această etapă s-a extras penultimul strat dens conectat şi s-a folosit pe post de descriptor la partea de regăsire.

Reţeaua a fost antrenată cu conceptul de marjă largă (Sectiunea 4.4) pentru a creşte densitatea spaţială a spaţiului descriptiv. Rezultatele sunt în tabelul 6.1. Ca metrici de performanţă s-au folosit precizia medie, eroarea şi aria de sub graficul curbei precizie-reamintire. Se poate observa că rezultatele sunt doar uşor în favoarea metodei prezentate.

Pentru a stabili utilitatea funcţiei de marjă largă s-au creat 3 noi scenarii care să conţină la nivel de descriptori date separabile, date neseparabile şi date din ambele categorii. O figură sugestivă cu spaţiul datelor înainte de antrenare se regăseste mai jos.

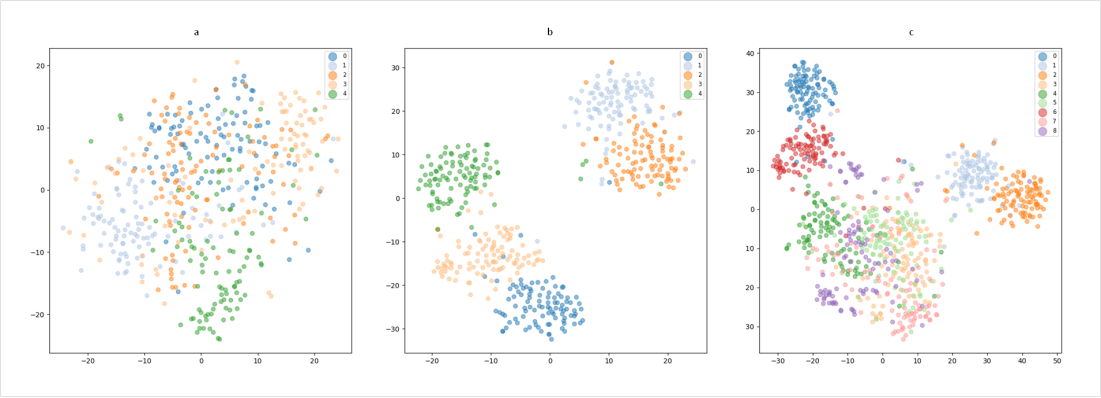

Figura 6.1 Reprezentarea spaţiului descriptiv pentru cele 3 scenarii. a – date neseparabile; b- date separabile; c- date separabile cu date neseparabile

Figură din [63]

În tabelul 6.2 sunt redate rezultatele pentru cele 3 sceanarii analizate. Ce trebuie remarcat este că funcţia de marjă largă conduce la rezultate mai bune atunci când spaţiul datelor este mai aglomerat. Pentru cazul cel mai defavorabil cu date foarte

Scenariu/Metrică [%]	Pr.med. 5-int.	Pr.med. 10-int.	Er. Top-1	Er. Top-5	Acc	ASG PR
EI-date separabile	92.82	92.93	6.85	2.46	94.82	82.81
ML-date separabile	92.82	92.93	6.85	2.46	94.82	**85.16**
EI-date neseparabile	55.83	56.00	43.61	12.69	67.20	37.25
ML-date neseparabile	59.75	58.97	40.60	13.45	70.08	**44.75**
EI-date separabile şi neseparabile	72.78	71.50	25.82	7.13	78.30	52.12
ML-date separabile şi neseparabile	72.11	72.43	22.15	9.32	79.85	**58.96**

Tabel 6.2 Rezultate experimentale Places 365 pentru noile scenarii (Pr.med.- precizie medie, Er. - Eroare, EI- entropie încrucişată , Acc. - acurateţe, ASG PR - aria de sub graficul precizie-reamintire, ML- marjă largă).

Tabel din [63]

greu separabile metricile de performanţă sunt clar mai bune decât în cazul utilizarii descriptorilor proveniţi din antrenarea cu entropia încrucişată.

Pentru a demonstra valoarea conceptului de marjă largă poate fi urmărită figura 6.2. Aici este expusă evoluţia spaţiului descriptiv pentru entropia încrucişată şi marja largă. Aşa cum se poate observa spaţiul este mult mai compact cu marja largă, ceea ce înseamnă că reţeaua va furniza descriptori mult mai utili pentru problema de regăsire de imagini.

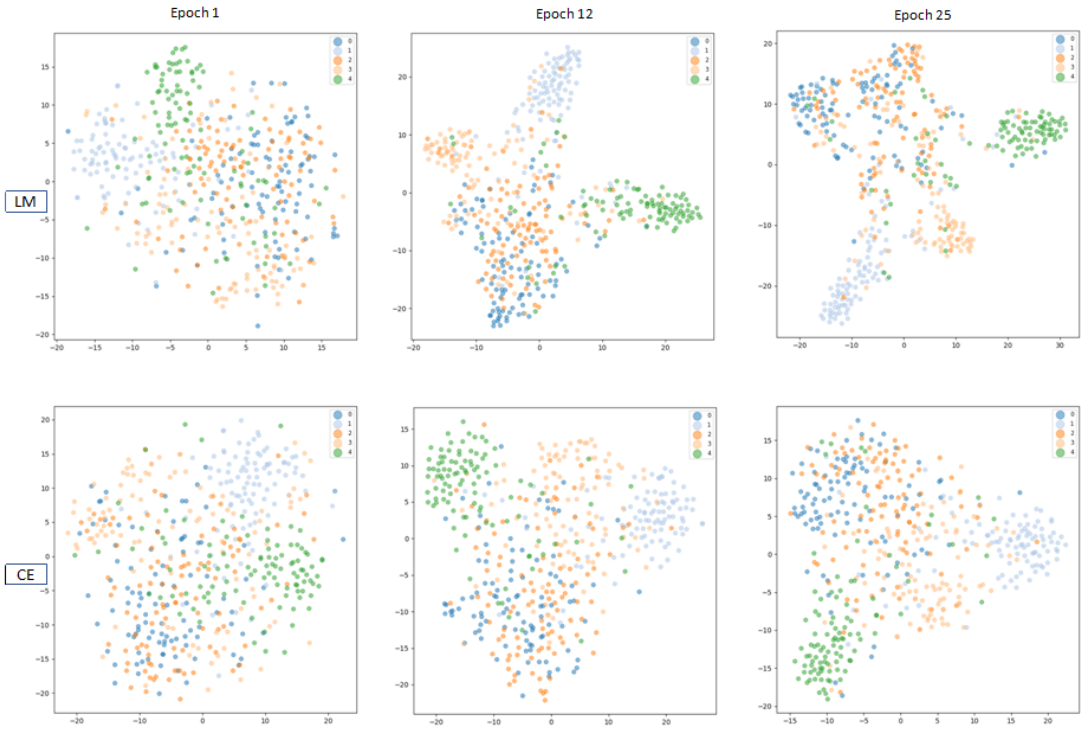

Figura 6.2 Evoluţia spaţiului descriptiv pe parcursul antrenării pentru cazul cu date suprapuse (ML –sus, EI- jos)
Figură din [63]

Ideea amintită anterior poate fi verificată în figura 6.3. Aici se pot vizualiza primele 5 imagini returnate pentru o serie de imagini de interogare. Se poate vedea că atunci când se folosesc descriptorii furnizaţi de reţeaua antrenată cu funcţia de marjă largă numărul de imagini returnate care se află în clasa corespunzătoare este mai mare. Cu toate acestea, există şi cazuri când entropia încrucişată oferă rezultate mai bune. Acest aspect demonstrează că pentru date foate suprapuse nici măcar conceptul de marjă largă nu e suficient.

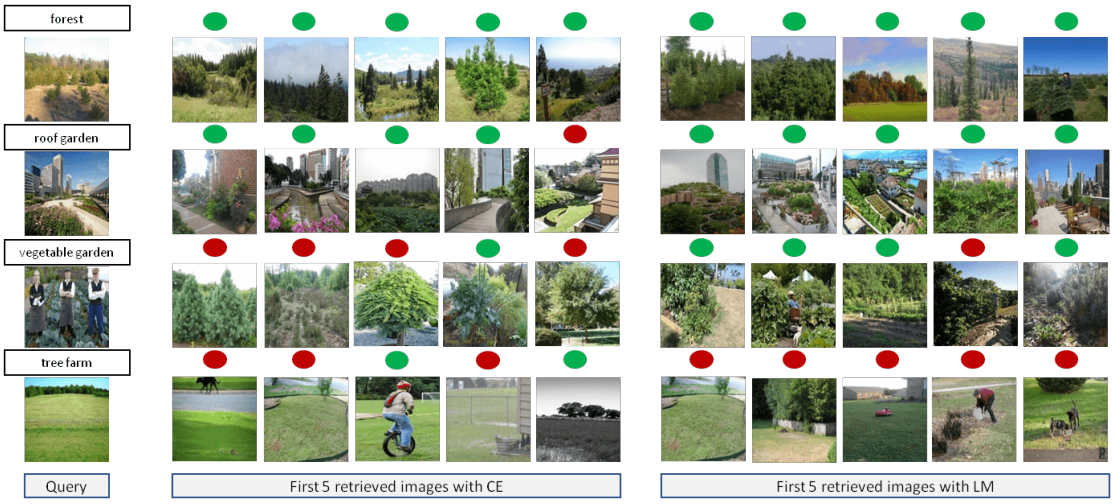

Figura 6.3 Exemple de primele 5 imagini returnate cu entropia încrucisată şi marja largă. Cercurile roşii reprezintă imagini returnate care au clasa diferită faţă de imaginea de interogare. Cercurile verzi reprezintă imagini returnate care au aceeaşi clasa ca imaginea de interogare
Figură din [63]

Chiar dacă marja largă are limitările sale, obţine o compactare mai bună a descriptorilor. În figura 6.4 este reprezentat la nivel descriptiv setul de referinţă din care sunt extrase imaginile pentru problema de regăsire. După cum se poate constata, atunci când se foloseşte rerţeaua antrenată cu marjă largă, datele sunt mai compacte si mai puţin suprapuse. Acest fapt uşurează deciziile sistemului în ceea ce priveşte regăsirea de poze cu conţinut similar. Creşterea ariei de sub graficul curbei precizie-reamintire din figura 6.5 confirmă utilitatea acestei tehnici pentru ptobleme cu date aglomerate.

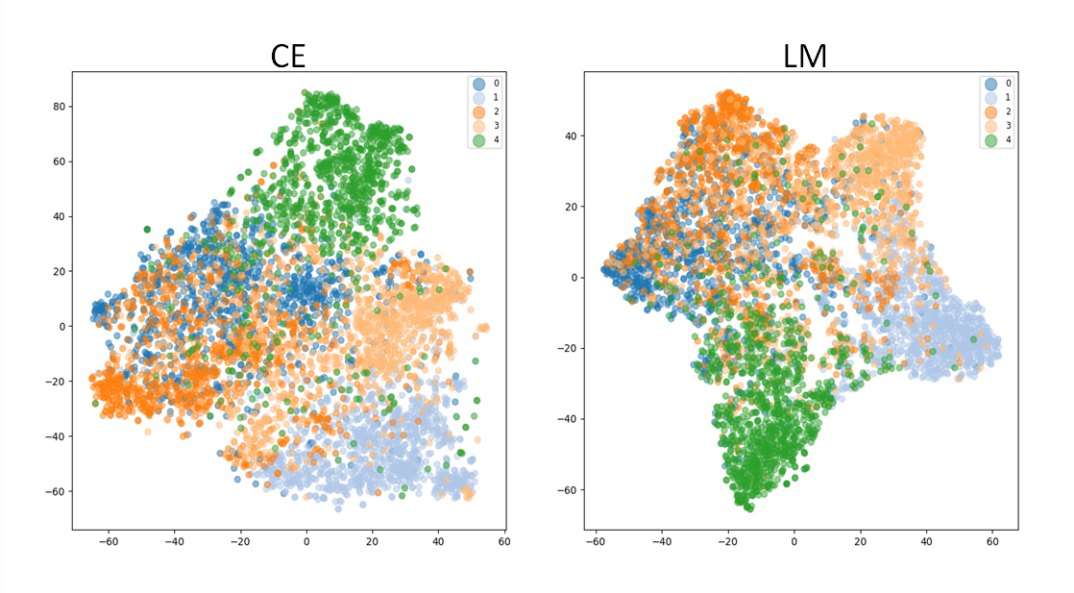

Figura 6.4 Reprezentarea descriptorilor pentru setul de referinţă pentru EI (stânga) şi ML (dreapta) în cazul cu date neseparabile
Figură din [63]

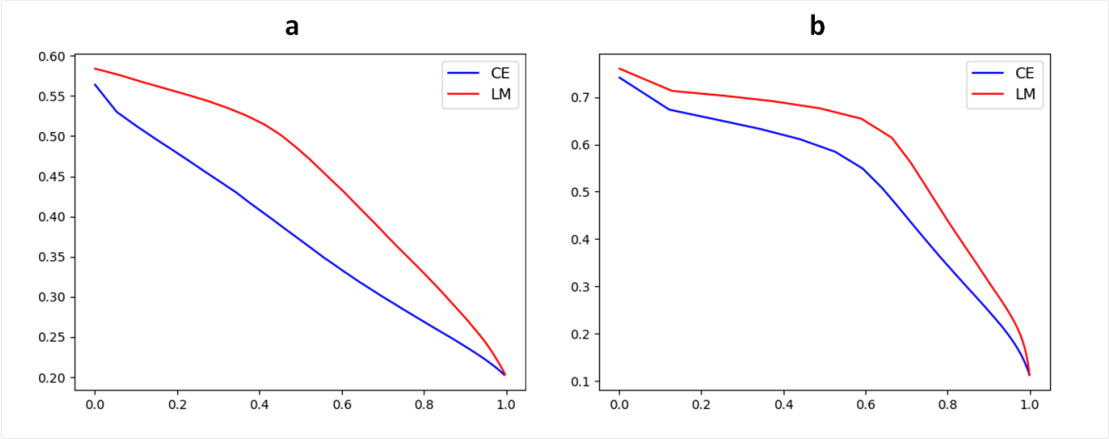

Figura 6.5 Curbele prefcizie-reamintire pentru EI şi ML. a- scenariul cu date neseparabile; b- scenariul cu date neseparabile şi date separabile)
Figură din from [63]

Capitolul 7

Concluzii

7.1 Rezultate obţinute

Pe parcursul lucrării au fost abordate 2 teme importante de cercetare: recunoaşterea expresiilor faciale şi regăsirea de imagini similare. Prima dintre ele s-a concentrat şi pe detecţia de mişcări faciale intitulate unităţi de acţiune. Au fost elaborate mai multe metode bazate pe învăţarea semi-supervizată/adaptare de domeniu, funcţii de cost cu rol de structurare al spaţiului descriptiv şi metode noi de augmentare/regularizare. Rezultatele satisfăcătoare au confirmat potenţialul metodelor propuse.

Celălalt domeniu de interes a fost regăsirea de imagini simale. În această situaţie s-a testat dacă o funcţie nouă de cost (marjă largă) poate grupa mai eficient descriptorii folosiţi mai departe pentru găsirea de imagini asemănătaore. S-a demonstrat că acest concept este util în special pentru cazurile în care datele se suprapun foarte mult la nivel descriptiv.

7.2 Contribuţii originale

- A fost abordată o nouă tehnică de recunoaștere a expresiilor faciale și a unităților de acțiune, care are mai multe componente originale. A fost folosită o soluție de adaptare a domeniului pentru a face legătura între mișcările faciale și expresiile discrete pentru a beneficia de potențialul învățării semi-supervizate. În plus, a fost utilizată o funcţie de cost cu rol de grupare a spațiului descriptiv pentru o performanță crescută. [14]

- A fost propusă o soluție pentru recunoașterea expresiilor faciale, concentrându-se pe utilizarea unei noi metode de regularizare bazată pe injectarea aleatore în gradient. [42, 43]

- A fost propusă o nouă modalitate care combină datele adnotate și datele neadnotate cu o tehnică de creștere a numărului de eşantioane. Algoritmul a fost testat atât în cazul expresiilor faciale, cât și pe baze de date standard [37]

- Eficacitatea unei funcţii de cost cu rol de grupare a spațiului descriptiv a fost testată pentru sarcina de regăsire de imaginii. În acest caz, o grupare mai eficientă a descriptorilor a crescut semnificativ performanța de regăsire pentru scenariile cu date mai suprapuse. [63]

- Metoda din [14] a fost extinsă pentru mai multe seturi de date care conţin unităţi de acţiune. Aici, funcția de pierdere propusă a fost studiată în raport cu alte funcții similare din literatura de specialitate și s-a dovedit a fi mai eficientă. De asemenea, s-a demonstrat că pierderea propusă contribuie la o mai bună recunoaștere a unităților de acțiune care apar mai rar în seturile de date, confirmând capacitatea acesteia de a grupa corespunzător datele. [48] [64]

- Pentru toți algoritmii propuși, a fost efectuată o comparație extinsă cu literatura de specialitate. Au fost discutate metode similare cu abordările studiate pentru a obține o idee mai obiectivă despre eficacitatea acestora.

7.2.1 Publicaţii

- Andrei Racoviteanu, Corneliu Florea, Mihai Badea, and Constantin Vertan. Spontaneous emotion detection by combined learned and fixed descriptors. In 2019 International Symposium on Signals, Circuits and Systems (ISSCS), pages 1–4. IEEE, 2019

- Andrei Racoviteanu, Iulian Felea, Laura Florea, Mihai Badea, and Corneliu Florea. Clustering based reference normal pose for improved expression recognition. In International Conference on Advanced Concepts for Intelligent Vision Systems, pages 51–61. Springer, 2018

- Mihai Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. UPB Scientific Bulletin, Series C: Electrical Engineering, pg Series C, Vol. 84, Iss. 3, 2022, pp 107-119

- Andrei Racoviteanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pages 1–4. IEEE, 2020

- M. Boeru, A. Racoviţeanu and C. Florea, "Facial Expressions Recognition by Structuring the Embeddings Space," 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2021, pp. 1-4
- B. Stoica, L. Florea, A. Bădeanu, A. Racoviţeanu, I. Felea and C. Florea, "Visual saliency analysis in paintings," 2017 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2017, pp. 1-4
- Badea, M., Florea, C., Racoviţeanu, A., Florea, L., Vertan, C. (2023). Timid semi–supervised learning for face expression analysis. Pattern Recognition, 138, 109417.
- Florea, Corneliu, et al. "Automatic Real-Estate Image Analysis for Retrieval and Classification." Bulletin of the Polytechnic Institute of Iași. Electrical Engineering, Power Engineering, Electronics Section 68.2 (2022): 35-45.
- Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, pages 1–17. Springer, 2020
- Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In BMVC, page 104, 2019
- Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In BMVC, page 108, 2019
- Andrei Racoviteanu, Corneliu Florea, Mihai-Sorin Badea. Large margin loss for Image Retrieval. Accepted to UPB Scientific Bulletin, Series C: Electrical Engineering
- Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Normalized Margin Loss for Action Unit Detection. Submitted to MVAP
- Project "Technologies and innovative video/audio systems for the recognition/identification of people and simulated behavior" - SPIA-VA, PN-III-P2-2.1-SOL-2016-02-0002
- Project "TRANSLATE" , TE 66/2020, PN-III-P1-1.1-TE-2019-0543.
- Project “Innovative Artificial Intelligence systems in the field of real estate portals” - online number 137-221-A2, MySMIS number: 129132
- Project "OPTIMizarea rezultatelor cercetării aplicative a doctoranzilor și cercetătorilor postdoctorat" - nr. 62461/03.06.2022, SMIS code 153735.

7.2.2 Dezvoltări ulterioare

Tehnica de adaptare a domeniului folosită pentru expresiile faciale poate fi folosită și în alte contexte. Localizarea punctelor faciale și clasificarea poziției capului sunt exemple practice, deoarece cele trei unghiuri ale capului pot fi corelate în raport cu punctele faciale; un alt exemplu este descrierea imaginilor și detectarea obiectelor, unde descrierile sunt derivate dintr-un anumit set de obiecte.

În cazul regăsării imaginilor, funcţia de cost cu rol de grupare a spațiului descriptiv s-a dovedit a fi foarte eficientă. În acest context, ar putea fi extinsă și la bazele de date cu expresii faciale. Având în vedere că multe expresii sunt similare la nivel descriptiv, această soluție are un anumit potențial.

Bibliografie

[1] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[7] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[8] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[9] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[11] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[12] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[13] Y. Zheng, D. Pal, and M Savvides. Ring loss: Convex feature normalization for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, 2018.

[14] Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In *BMVC*, page 108, 2019.

[15] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[16] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[17] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[18] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015.

[19] Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 108–116, 2018.

[20] Zixing Zhang, Jing Han, Jun Deng, Xinzhou Xu, Fabien Ringeval, and Björn Schuller. Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. *IEEE Access*, 6:22196–22209, 2018.

[21] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 298–313, 2018.

[22] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[23] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017.

[24] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28, 2015.

[25] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[26] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[27] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[28] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[29] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

[30] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016.

[31] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[32] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[33] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:1532, 2015.

[34] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer, and Saida Bouakaz. A novel database of children's spontaneous facial expressions (liris-cse). *Image and Vision Computing*, 83:61–69, 2019.

[35] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[36] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[37] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 1–17. Springer, 2020.

[38] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[39] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[40] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[41] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

[42] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019.

[43] Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Randomization injection for efficient transfer in face expression recognition. In *Submitted to Applied Intelligence*, 2023.

[44] Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 589–593, 2018.

[45] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[46] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[47] Yanling Gan, Jingying Chen, and Luhui Xu. Facial expression recognition boosted by soft label with a diverse ensemble. *Pattern Recognition Letters*, 125:105–112, 2019.

[48] Andrei Racoviteanu, Corneliu Florea, Laura Florea, and Constantin Vertan. Normalize margin loss for action unists detection. In *Submitted to MVAP*, 2023.

[49] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[50] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6766–6775, 2017.

[51] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[52] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[53] Jiyuan Cao, Zhilei Liu, and Yong Zhang. Cross-subject action unit detection with meta learning and transformer-based relation modeling. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[54] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5993–6001, 2021.

[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[56] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[57] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[58] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[59] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006.

[60] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[61] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing*, 9(5):489, 2017.

[62] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[63] Andrei Racoviteanu, Corneliu Florea, and Mihai Badea. Large margin loss for image retrieval. In *Accepted to UPB Scientific Bulletin, Series C: Electrical Engineering*, 2023.

[64] Andrei Racovițeanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2020.