MINISTERUL EDUCAȚIEI NAȚIONALE

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Sistem multifocal pentru urmărirea în timp real a trăsăturilor dinamice faciale și corporale (MULTIFACE)

Raport sintetic proiect.
Anii: 2015 - 2017
Director proiect: Prof. Dr. Ing. Radu Dănescu
Echipa de cercetare: Radu Dănescu, Florin Oniga, Diana Borza, Razvan Itu, Mircea Muresan

Cuprins

1. Rezumatul proiectului

În cadrul acestui proiect de cercetare s-au dezvoltat algoritmi, metode şi sisteme pentru urmărirea non-obtruzivă a capului şi a elementelor faciale, fără impunerea de constrângeri de mişcare şi de expresie, în vederea determinării diferiților parametri ce pot fi folosiți în biometrie, interacțiunea om-maşină, şi determinarea stărilor psihologice.

Inițial am efectuat o analiză comprehensivă a algoritmilor de stereoviziune existenți pentru măsurători ale capului şi ale trăsăturilor faciale, iar algoritmi de stereoviziune au fost îmbunătățiți pentru monitorizarea capului şi a feței.

Urmărirea capului şi analiza macro-expresiilor este un domeniu al viziunii computerizate care a fost intens studiat în ultima perioadă şi această problemă poate fi considerată rezolvată: deja au apărut mai mulți algoritmi *open-source* de detecție a poziției capului şi de detecție a trăsăturilor faciale. Astfel, deşi în propunerea de proiect, am propus un obiectiv cu urmărirea şi localizarea trăsăturilor faciale, am decis că acest obiectiv nu mai este de actualitate şi ne-am focalizat pe alte activități care au fost mai puțin studiate: urmărirea ochilor în condiții neconstrânse (*unconstrained*), detecția şi recunoaşterea micro-expresiilor şi analiza trăsăturilor faciale.

Ochii sunt cele mai importante elemente ale feței, iar mişcările lor au un rol important în exprimarea stărilor emoționale şi a proceselor cognitive. În cadrul acestui proiect s-au propus trei metode originale pentru urmărirea ochilor şi segmentarea acestora : găsirea centrului şi a razei irisului şi segmentarea pleopapelor/sclerei. Pentru a găsi centului irisului toate metodele propuse utilizează un detector de simetrii circulare: *Fast Radial Symmetry Transfrom FRST* [16]. *FRST* este o transformare a imaginii care utilizează gradientul imaginii pentru a determina rolul pe care fiecare pixel *p* îl are la simetria pixelilor vecini aflați la o distanță *r* de acest pixel. Această contribuție se calculează prin însumarea contribuțiilor magnitudinii şi orientării pe direcția gradientului. Candidații pentru centrelor irişilor sunt determinați ca minime locale din transformata *FRST,* iar centrele irisurilor sunt selectate din candidații determinați la pasul anterior pe baza unor constrângeri geometrice. Raza irisurilor se calculează utilizând derivata Sobel a imaginii pe o regiune în jurul ochilor pentru a accentua tranziția puternică dintre zona irisului şi a sclerei.

Pentru segmentarea formei externe a ochilor, am reprezentat forma ochilor prin două parabole, una pentru pleoapa de sus şi una pentru pleoapa de jos. Pentru potrivirea unei forme ipotetice a ochilor la o imagine, am propus mai multe metode. Prima metodă se bazează pe informații de culoare : regiunea din jurul ochiului este transformată într-un spațiu probabilistic utilizând algoritmi de învățare automată pentru a determina probabilitatea unui pixel de a fi un pixel din regiunea sclerei. Apoi această metodă a fost rafinată şi s-au mai multe informații, nu doar informații despre culoare, pentru măsurarea gradului în care o particulă se potriveşte în imaginea de intrare : colțuri, transformata distanță şi culoare.

Următorul pas a fost dezvoltarea unei metode care să permită urmărirea ochilor în cadrele video. Soluția propusă foloseşte o abordare de la grosier la fin (*coarse-to-fine*) pentru a detecta şi urmări multiple trăsături ale ochilor: centrul irisului, conturul ochiului şi clipirile. Soluția utilizează trei filtre de particule în paralel pentru a urmări ochii: primul filtru de particule este folosit pentru a determina pozițiile aproximative ale irisurilor. Alte două filtre de particule sunt utilizate pentru a determina şi a urmări conturul fiecărui ochi pe baza estimării obținute de la primul filtru de particule.

În prima etapă se utilizează un filtru de particule pentru a estima la modul grosier poziția centrilor irisurilor şi orientarea ochilor (unghiul dintre linia care unește cei doi ochi şi axa

orizontală). În primul cadru video, filtrul de particule este inițializat: fața este detectată în imaginea de intrare și spațiul de căutare este în mod uniform acoperit prin generarea aleatoare de particule în întregul spațiu de intrare.

În cadrele video următoare, se aplică iterativ filtrul de particule pentru a actualiza estimarea poziției ochilor. În mod periodic, din *t* în *t* cadre, estimarea obținută este supusă unui test de validitate: fața este detectată în cadrul curent și se verifică dacă pozițiile estimate ale ochilor se află în zona aproximativă a ochilor (regiunea din jumătatea de sus a feței). Dacă estimarea nu mai este validă, atunci urmărirea este re-inițializată.

După analiza mișcărilor ochilor, ne-am focalizat pe analiza micro-expresiilor. Micro-expresiile sunt expresii faciale scurte, cu o durată cuprinsă între 1/12 și 1/25 secunde, și apar atunci când oamenii încearcă să-și ascundă sentimentele, fie ca o formă de surpresie (ascundere intenționata), fie ca o formă de refulare (ascundere inconștientă). Deși recunoașterea automată a (macro) expresiilor a fost intens studiată în ultimii ani, analiza automată a micro-expresiilor este un subiect relativ nou și puține studii s-a efectuat în această direcție.

La început, am propus o metodă de detecție și recunoaștere a micro-expresiilor bazată pe rețele neuronale convoluționale (*convolutional neural networks*). Rețeaua primește ca intrare două imagini diferență. De asemenea, am propus și o metodă de post-procesare al răspunsului rețelei care îmbunătățește considerabil performanța sistemului.

Ulterior am dezvoltat o metodă care analizează magnitudinea mișcării apărute de-a lungul secvenței video pentru a determina momentele în care apar micro-expresiile, precum și momentele cheie ale acestora: punctul de *onset*, punctul de *apex* și punctul de *offset*. Metoda propusă extrage trăsături de mișcare din anumite zone de pe față care corespund mușchilor faciali implicați în exprimarea emoțiilor și utilizează un clasificator *Random Forest Classifier* pentru a determina momentele în care apar micro-expresiile.

În final, această metodă a fost rafinată și extinsă pentru a recunoaște și tipul micro-expresiei (de tip pozitiv, de tip negativ sau de tip surpriză). Pentru aceasta s-a propus un descriptor original de mișcare bazat pe deplasamentul relativ al centrelor regiunilor faciale corespunzătoare mușchilor faciali. Metoda de recunoaștere nu necesită antrenare și are performanțe superioare metodelor propuse în literatura de specialitate.

Analiza trăsăturilor facile: rasă, sex, vârstă este un domeniu care începe să fie din ce în ce mai studiat în literatura de specialitate. În cadrul acestui proiect ne-am focalizat pe analiza rasei/etniei și a sexului. Am propus două metode bazate pe rețele neuronale convoluționale pentru recunoașterea acestor atribute. O altă contribuție importantă este faptul că pentru fiecare metodă (sex și rasă), am colectat câte o bază de date de mari dimensiuni (60000 imagini, respectiv 200000 imagini) cu imagini faciale pe care le-am făcut publice.

Algoritmii de urmărire și analiză a expresiilor au fost integrați într-o aplicație demonstrator pentru monitorizarea și analiza răspunsului emoțional al oamenilor. Sistemul poate fi utilizat fie pentru a capta și a stoca secvențele video de înaltă viteză, fie pentru a analiza online (pentru a detecta și a recunoaște micro-expresii sau pentru a urmări poziția ochilor) *stream*-ul video.

Rezultatele proiectului sunt relevante pentru toate domeniile științifice interesate în observarea detaliată a indivizilor: psihologie, psihiatrie, educație, monitorizarea atenției șoferilor, biometrică, etc.

2. Rezumatul etapelor

2.1. Etapa 1

Această etapă a avut o durată de timp foarte limitată, astfel încât activitățile prevăzute au fost împărțite în sub-activități, fiind realizate doar sub-activitățile preliminare în prima etapă, celelalte fiind alocate etapei 2.

Mai precis, în această etapă s-au demarat activitățile în vederea realizării obiectivelor proiectului: realizarea şi calibrarea sistemului multifocal, realizarea sistemului de stereoviziune, şi modelarea şi urmărirea capului si a trăsăturilor faciale.

A fost realizat un studiu amplu în vederea realizării sistemului senzorial multifocal, şi au fost făcute primele achiziții, conținând componente de mica valoare, urmând ca achiziția componentelor principale sa fie făcuta in etapa a doua.

A fost definită de asemenea metodologia de calibrare, si au fost studiate performantele a multipli algoritmi de stereoviziune. A fost făcut de asemenea un studiu detaliat privind modelarea si urmărirea trăsăturilor faciale, definind şi potențialele contribuții originale punctuale care pot fi aduse.

În limita bugetului din 2015, şi a timpului scurt pentru implementare acestei etape, au fost efectuate doar achiziții de mică valoare, urmând ca cele care implică un cost mai mare, şi implicit o procedură de durată, să fie efectuate în etapa următoare.

În vederea constituirii echipei independente de cercetare, în aceasta etapă au fost angajați doi cercetători doctoranzi care nu au mai avut contract de muncă cu Universitatea Tehnica, şi un cercetător masterand.

2.2. Etapa 2

Pentru îmbunătățirea calității reconstrucției tridimensionale, au fost proiectate şi implementate metode noi pentru calculul corespondențelor stereo, care au reuşit să producă rezultate cu densitate superioară, fără a folosi însă metode de optimizare globală.

S-a stabilit un model 3D pentru fața umană care va fi utilizat pentru a urmări trăsăturile faciale şi pentru a calcula orientarea capului. Pentru înțelegerea modului în care ar trebui potrivit modelul 3D peste imagini, s-a dezvoltat o aplicație care permite suprapunerea manuală a modelului peste o imagine. Pe baza observațiilor obținute, s-a proiectat şi s-a implementat un algoritm care potrivește automat şi rapid modelul peste o imagine, determinându-se astfel orientarea capului şi localizarea elementelor feței (ochi, buze, sprâncene).

De asemenea, am dezvoltat 2 metode originale de detecție a trăsăturilor ochilor (centrul irisului, colțurile ochilor şi pleoapele) în imagini şi o metodă originală de urmărire a acestor trăsături şi de detecție a clipirilor în cadre video.

Pentru realizarea acestui sistem, au fost studiate mai multe opțiuni pentru camere video, lentile, arhitectura de procesare şi mecanismele de control şi orientare a camerelor. S-a optat pentru un sistem compus din două camere alb-negru, fixe, pentru stereoviziune, montate împreună cu o a treia cameră, care va fi color, şi va avea capacitatea de achiziție rapidă (mai mult de 200 de cadre pe secundă).

În această etapă s-a realizat achiziția principală a acestui proiect: două camere de viteză mare (*high speed*) XIMEA. În plus s-au mai achiziționat anumite componente mecanice necesare realizării prototipului unității *pan tilt:*

- Motoare pas cu pas

- Plăci SOC de tip Raspberry PI 2
- Plăci cu microcontroller Arduino Mega 2560
- Drivere motor de tip Arduino Motor Shield
- Cabluri, surse de alimentare, alte accesorii pentru montaj.

2.3. Etapa 3

În această etapă au fost continuate activitățile începute în anul anterior în vederea realizării obiectivelor proiectului:

- testarea și evaluarea algoritmilor de urmărire a capului și a trăsăturilor faciale
- proiectarea, implementarea și evaluarea unor algoritmi pentru detecția și recunoașterea micro-expresiilor
- proiectarea, implementarea și evaluarea unor algoritmi pentru recunoașterea atributelor faciale.

În primul rând am testat și evaluat mai multe sisteme existente (*off the shelf*) de urmărire a poziției capului și a trăsăturilor faciale pentru a determina dacă acestea pot fi utilizate în sistemul propus. În urma analizei efectuate am decis că librăria C++ FaceAnalysis SDK corespunde cerințelor sistemului și am integrat-o în aplicația dezvoltată.

În continuare ne-am focalizat pe detecția și recunoașterea micro-expresiilor în secvențe video *high speed*. Concret, am propus și dezvoltat trei metode originale de detecție/recunoaștere a micro-expresiilor.

Prima metodă propusă se bazează pe rețele neuronale convoluționale (*convolutional neural networks*) și propune o taxonomie cu patru clase: non-microexpresie, micro-expresie de tipul surpriză, micro-expresie negativă și micro-expresie pozitivă. Rețeaua primește ca intrare două imagini diferență. De asemenea, am propus și o metodă de post-procesare al răspunsului rețelei care îmbunătățește considerabil performanța sistemului.

A doua metodă propusă analizează magnitudinea mișcării apărute de-a lungul secvenței video pentru a determina momentele în care apar micro-expresiile, precum și momentele cheie ale acestora: punctul de *onset*, punctul de *apex* și punctul de *offset*. Metoda propusă extrage trăsături de mișcare din anumite zone de pe față care corespund mușchilor faciali implicați în exprimarea emoțiilor și utilizează un clasificator *Random Forest Classifier* pentru a determina momentele în care apar micro-expresiile.

În final, această metodă a fost rafinată și extinsă pentru a recunoaște și tipul micro-expresiei (de tip pozitiv, de tip negativ sau de tip surpriză). Pentru aceasta s-a propus un descriptor original de mișcare bazat pe deplasamentul relativ al centrelor regiunilor faciale corespunzătoare mușchilor faciali. Metoda de recunoaștere nu necesită antrenare și are performanțe superioare metodelor propuse în literatura de specialitate.

În ceea ce privește recunoașterea atributelor faciale, am extras trei tipuri de atribute: genul, culoarea pielii și rasa/etnia. Algoritmii propuși au fost proiectați și implementați pentru un sistem de visagism. Visagismul este un concept nou apărut în industria modei și a optometriei care urmărește să pună în evidență sau să atenueze anumite trăsături faciale prin utilizarea unor accesorii care să fie în armonie cu fața clientului. În plus, metodele propuse au numeroase alte aplicații, cum ar fi: interacțiune om calculator, securitate, medicină etc.

În final, am dezvoltat o aplicație demonstrator în care am integrat toți algoritmii implementați în cadrul acestui proiect. Aplicația are două regimuri principale de funcționare: (1) Detecția și recunoașterea în timp real a reacțiilor emoționale ale utilizatorului. (2)

Inducerea emoțiilor (prin afișarea unor stimuli video) și captura video la o rezoluție temporală mare a reacției utilizatorului.

3. Prezentarea detaliată a etapei finale

3.1. Testarea și validarea algoritmilor de urmărire a capului și a trăsăturilor faciale

Metodele de urmărire a capului și a trăsăturilor faciale constituie un domeniu de cercetare activ, cu implicații într-o multitudine de domenii pluridisciplinare. În cadrul acestei activități s-au evaluat mai multe metode *open-source* pentru urmărirea capului și a trăsăturilor faciale pentru a determina stadiul actual în domeniu și eventualele îmbunătățiri care pot fi aduse.

Inițial au fost stabilite următoarele specificații minime care ar trebui să fie îndeplinite de modul pentru urmărirea capului și a trăsăturilor faciale. În primul rând, sistemul de urmărire a poziției capului ar trebui să fie autonom: nu ar trebui să fie necesară inițializarea manuală a sistemului. În plus, sistemele de urmărire a poziției capului ar trebui să fie capabile să estimeze continuu poziția capului în timp real, atât în imagini de apropiere, cât și în imagini de depărtare. Sistemele ar trebui să permită o gamă largă de mișcări (să funcționeze chiar și atunci când fața nu este îndreptată spre cameră) și să furnizeze estimări precise.

S-au evaluat următoarele metode/sisteme:

- *Face Analysis SDK* [1, 2] este o bibliotecă C++ care implementează mai mulți algoritmi pentru extragerea și analiza geometriei capului. Algoritmul găsește 66 de puncte pe fețele umane prin potrivirea unui model 3D la pixelii din imagine folosind o versiune îmbunătățită a metodei *Deformable Model Fitting by Regularized Landmark Mean Shift* [2]. Pentru fiecare din cele 66 de puncte se returnează și poziția 3D a punctului. În plus, pentru a crește robustețea sistemului, biblioteca implementează și un modul de detecție al erorilor de urmărire.
- *Dlib* [3] – biblioteca C++ propune un modul de detecție al fetelor umane (în poziție frontală) și estimează poziția acestor pe baza a 68 de puncte de față (colțurile ochilor, ale gurii, sprâncenele etc). Se utilizează un detector de fețe pe baza trăsăturilor de tipul Histogram of Oriented Gradients (HOG) combinate cu un clasificator linear, imagini piramidă și o metodă de detecție ce utilizează un *sliding window*. Estimarea poziției capului se face în timp real.
- *Kinect Face Tracking* [4]: framework-ul Face Tracking SDK dezvoltat de Microsoft permite urmărirea fețelor umane în timp real folosind un dispozitiv Kinect. Framework-ul analizează cadrele furnizate de dispozitivul Kinect și determină poziția și orientarea capului, localizează 87 de puncte pe față și determină emoțiile subiecților din cadrele video.
- *Intel Real-Sense* [5] propune mai multe soluții pentru sistemele de interacțiune om-calculator bazate pe gesturi. Metodele propuse de Intel utilizează camere 3D și o librărie software de *machine perception* pentru a localiza 68 de puncte pe față și pentru a estima poziția capului.

În urma evaluării efectuate s-a determinat că biblioteca C++ Face Tracking SDK corespunde întru totul cu cerințele formulate. În plus, biblioteca este implementată în C++, deci poate fi cu ușurință portată pe diferite sisteme de operare (chiar și pe mobile) și nu utilizează alte framework-uri (cum e biblioteca *boost* utilizată de *dlib*).

3.2. Proiectarea algoritmilor pentru detecția și recunoașterea micro-expresiilor

Recunoașterea automată a expresiilor faciale este un domeniu care a fost intens studiat în ultimii 30 de ani în domeniul viziunii artificiale, și deja există în industrie mai multe aplicații pentru detectarea și analiza macro-expresiilor din secvențe video.

Deși micro-expresiile sunt din ce în ce mai studiate pentru a înțelegerea comportamentul uman, ele au niște caracteristici care fac foarte dificilă detecția lor automată.

În primul rând, acestea sunt mișcări involuntare, deci este dificil să se obțină datele de test. La ora actuală sunt disponibile doar 4 baze de date cu micro-expresii: USF-HD, SMIC, CASME și CASME II. Dintre acestea doar SMIC [6], CASME [7] și CASME II [8] conțin expresii spontane. Baza de date SMIC conține secvențe video capturate cu o camera de mare viteza de 100 FPS:164 de micro-expresii de la 16 participanți; micro-expresiile sunt clasificate în trei categorii (pozitive, negative și surpriză). Baza de date CASME a fost capturată cu o cameră la o frecvență de 60 FPS și conține 195 de expresii spontane ale celor 20 de participanți, iar expresiile au fost adnotate în 7 categorii, pe baza metodologiei FACS. CASME II este o versiune ulterioară a lui CASME care conține 247 de micro-expresii capturate în condiții de iluminare controlată, cu o camera de mare viteză de 200 FPS. În plus, la ora actuală, CASME II este baza de date cu cea mai mare rezoluție a feței.

A doua problemă este faptul că ele sunt vizibile doar un număr redus de cadre, iar intensitatea mișcărilor faciale apărute în micro-expresii este foarte redusă. Detecția micro-expresiilor necesită deci algoritmi precișI de urmărire și detecție a mișcării.

În cadrul acestei etape s-a propus un sistem complet de analiză a micro-expresiilor ce folosește descriptori de mișcare originali bazați pe simpla diferență dintre cadrele succesive în videoclipul de intrare. Un sistem complet de analiză al micro-expresiilor (Figura 1) trebuie să conțină două module principale: un modul de detecție (care determină când apare o anumită micro-expresie) și un modul de recunoaștere (care determină tipul micro-expresiei apărute).

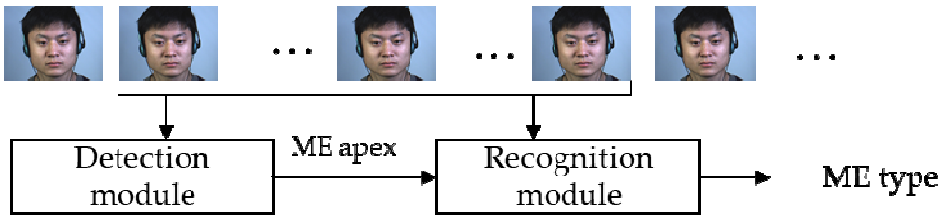

Figura 1. Sistemul propus pentru analiza micro-expresiilor

Modulul de detecție se bazează pe amplitudinea mișcării care survine de-a lungul cadrelor video de înaltă viteză, calculată prin intermediul unor diferențe absolute simple între imagini. Informațiile cu privire la mișcare sunt extrase din fiecare cadru și s-a folosit algoritmul Adaboost [10, 11] pentru a decide dacă un cadru aparține clasei „micro-expresie” (ME) sau clasei „non-micro-expresie”. Structura modulului de detecție este prezentată în Figura 2.

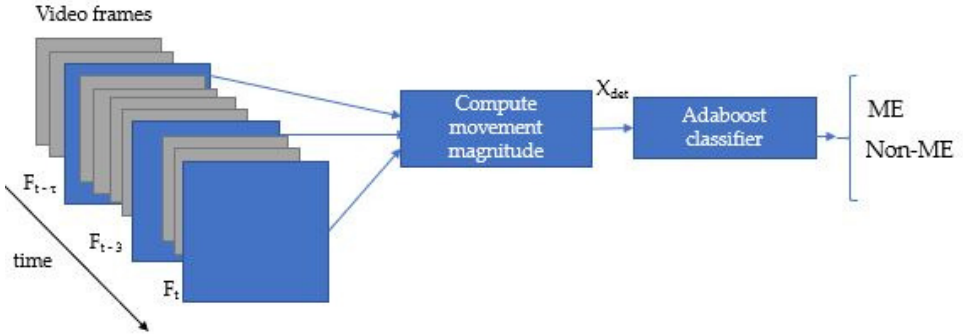

Figura 2. Modulul pentru detecția micro-expresiilor

Să notăm cu τ durata medie a unei micro-expresii în număr de cadre; am calculat această valoare ca fiind 67 pentru setul de date CASME II și 37 pentru setul de date SMIC-E.

Întrucât scopul modulului de detecție este de a găsi cadrele *apex*, considerăm diferența imagistică absolută dintre cadrul curent t (un cadru *apex* potențial) și cadrul anterior de la o distanță $\tau/2$ (un cadru *onset* potențial). Cu toate acestea, întrucât mișcările faciale care survin în timpul unei micro-expresii au o intensitate foarte scăzută, am introdus și un factor de normalizare pentru a distinge mișcarea de tip ME de zgomotul cauzat de condițiile de iluminare sau de dispozitivele de captare. Cadrul t - ε (ε = 3 în experimentele noastre) este folosit ca factor de normalizare. Deoarece secvențele video sunt capturate cu camere de înaltă viteză, nici o mișcare facială nu ar trebui să se producă în 0.03 s (valoare calculată pentru o rezoluție temporală de 100 cadre pe secundă).

În final, variația mărimii mișcării este calculată ca diferența absolută dintre cadrul t și t - $\tau/2$ normalizată cu diferența absolută dintre frame-ul curent și frame-ul t-ε (Ecuația (1)).

(1)

Figura 2 prezintă imaginile obținute prin diferență între cadre care sunt folosite pentru task-ul de detecție a micro-expresiilor.

Cu toate acestea, numai 10 regiuni ale feței sunt analizate. De aceea, pentru fiecare celulă c, se calculează valoarea medie a imaginii magnitudinii mișcării ( )) în interiorul acelei regiuni de interes (ROI – *region of interest*). De exemplu, Figura 3 prezintă variația lui pentru celula mediană de la nivelul sprâncenelor pe durata unei secvențe de micro-expresie.

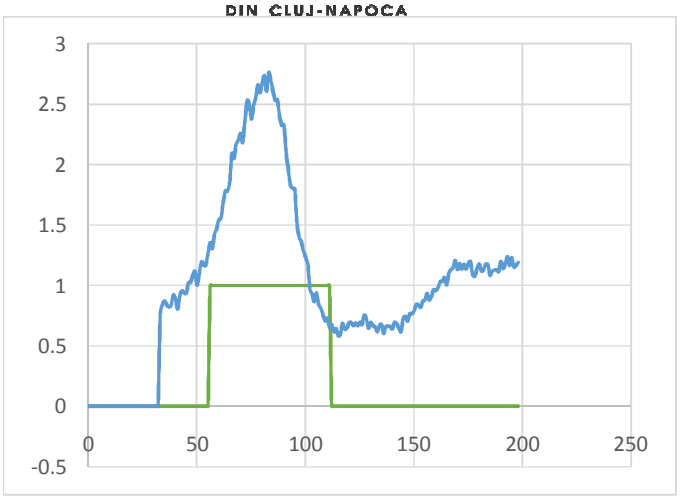

Figura 3. Variația în celula mediană de la nivelul sprâncenelor de-a lungul unei secvențe ME. Eticheta „ground truth" a secvenței ME este marcată cu verde pe imagine, iar valoarea lui este reprezentată cu albastru. Primele $\tau/2$ sunt ignorate (este setat la zero) deoarece imaginea MM nu poate fi calculată pentru cadrele t, unde $t < \tau/2$. Pe axa orizontală este reprezentat indexul cadrului din secvența ME, iar pe axa verticală este reprezentată valoarea medie a variației magnitudinii mișcării în interiorul acestei celule.

Pentru a extrage vectorul de trăsături în vederea clasificării unui cadru se iau în considerare toate celulele. Pentru un nou cadru de test i, o fereastră de dimensiune τ este centrată în cadrul curent. Pentru fiecare dintre cele 10 celule faciale (c) se extrage valoarea minimă și maximă a lui n intervalul $[i - \tau/2, i + \tau/2]$. Cu alte cuvinte, vectorul de trăsături extras pentru fiecare cadru poate fi exprimat ca:

= [min( ), max( ), min( ), max( ) , …., min( ), max( )] (2)

, unde reprezintă variația valorilor pentru celula din fereastra temporală centrată în cadrul curent:

(3)

Pentru etichetarea imaginilor de antrenament în micro și non-micro cadre se folosește următoarea regulă:

Dacă $t \in [0, t_{apex} - \delta \cdot \tau]$ sau $t \in [t_{apex} + \delta \cdot \tau, seqLen]$, atunci cadrul t este etichetat drept cadru non-micro-expresie (cadru neutru sau macro-expresie), atunci când *seqLen* este lungimea video-secvenței, în cadre.

Dacă $t \in (t_{apex} - \delta \cdot \tau, t_{apex} + \delta \cdot \tau)$, atunci cadrul t este considerat a fi un cadru ME

, unde δ este setat la 0.25. Cu alte cuvinte, definim un interval de dimensiune egală cu jumătate din micro-expresia medie, centrat în cadrul *apex*, care va conține cadrele etichetate drept cadre ME.

Totuși, setul de antrenament este foarte dezechilibrat: există mult mai multe cadre non-ME decât cadre ME. Pentru antrenament folosim toate cadrele ME disponibile și selectăm în mod aleatoriu un număr egal de cadre non-ME.

În final, vectorul de trăsături este furnizat ca intrare algoritmului Adaboost pentru a determina tipul fiecărui cadru. Adaboost este un clasificator meta-estimator care folosește un

set de algoritmi de învățare (sau estimatori) „slabi" care sunt combinați într-o sumă ponderată în vederea îmbunătățirii performanței clasificării. La fiecare iterație a algoritmului de învățare, estimatorii slabi sunt ajustați astfel încât să se focalizeze pe instanțele clasificate greșit de către algoritm la pașii anteriori. Am folosit 35 de estimatori slabi (*Decision Tree Classifiers*).

Folosind algoritmul și etichetarea descrise mai sus, ne-am putea aștepta ca clasificatorul să prezică multiple micro-cadre în jurul *apex*-ului real. De aceea, răspunsul clasificatorului este în continuare post-procesat în vederea filtrării soluțiilor fals pozitive și pentru a fuziona răspunsurile pozitive care aparțin aceleiași ME. Mai întâi toate intervalele ME disjuncte sunt detectate și intervalele care sunt prea apropiate unul de altul sunt fuzionate. În final, dimensiunea fiecărui interval este examinată, iar intervalele care sunt prea scurte sunt eliminate (Algoritmul 1). Cadrele *apex* sunt setate la mijlocul fiecărui interval prezis a conține o ME.

Arhitectura modulului de recunoaștere al micro-expresiilor este ilustrat în Figura 4.

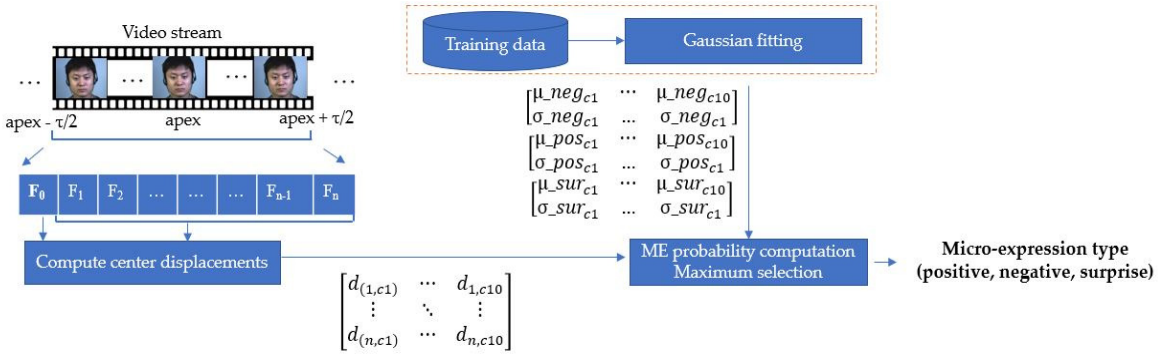

Figura 4. Modulul de recunoaștere a micro-expresiilor

O fereastră temporală (*sliding window*) de dimensiune τ este centrată în frame-ul cu *apex* detectat de modulul de detecție a micro-expresiilor și din interiorul acestei ferestre se extrag informații legate de traiectoria mișcării pentru a recunoaște tipul micro-expresiei. Pentru a asigura invarianța algoritmului față de frame-rate-ul camerei, acest interval temporal este discretizat în $n = 11$ imagini $\{F_0, F_1, F_2, \ldots, F_n\}$ (Figura 5).

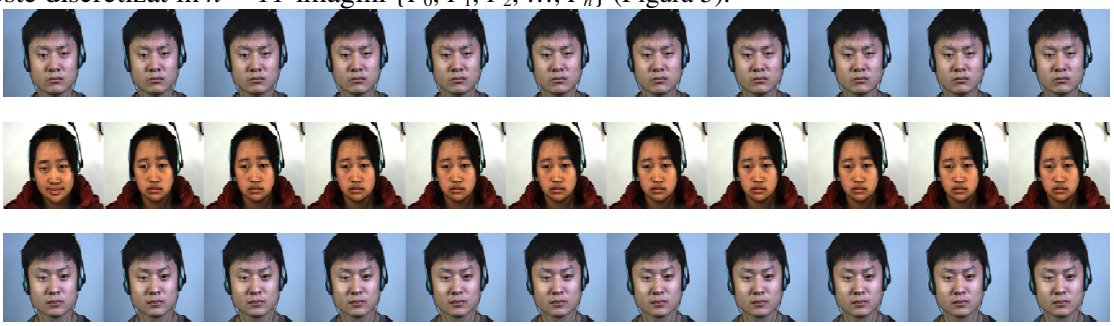

Figure 5. Exemple ale celor 11 imagini utilizate pentru a recunoaște tipul micro-expresiei. Pe primul rând micro-expresie negativă, pe al doilea rând micro-expresie pozitivă și pe cel de-al treilea rând un exemplu de micro-expresie de tip surpriză (imagini din baza de date CASME II (©Xiaolan Fu)).

Pentru a extrage informația legată de mișcare am propus un descriptor simplu bazat pe imaginea *movement magnitude (MM)*. Pentru fiecare celulă se calculează poziția ponderată centroidului pe baza intensității fiecărui pixel din imaginea *movement magnitude (MM):*

(4)

(5)

, unde reporezintă suma pixelilor din imaginea *MM* dintr-o celulă, *MM(x, y)* este valoarea pixelului din imaginea MM de la poziția (*x*, *y*) iar c_s, r_s, c_M, r_M definesc regiunea unei celule (*bounding rectangle*).

Poziția centrului (*cx0*, *cy0*) primei imagini F_0 este considerată poziția de referință pentru cazul neutru. În continuare se calculează diferențele dintre pozițiile ponderate ale centroizilor fiecărei imagini F_i, și poziția de referință: (cx_0, cy_0) . Aceste deplasamente constituie vectorul de trăsături pentru algoritmul de recunoaștere al micro-expresiilor:

(6)

Pentru etapa de antrenare, vectorul de trăsături X_i este extras pentru fiecare imagine din secvență și o funcție Gaussiană bidimensională este potrivită peste aceste date (Figura 6). Cu alte cuvinte, pentru fiecare tip de micro-expresie calculăm media și covarianța deplasamentelor relative ale centrelor pentru fiecare celulă: , , , , , .

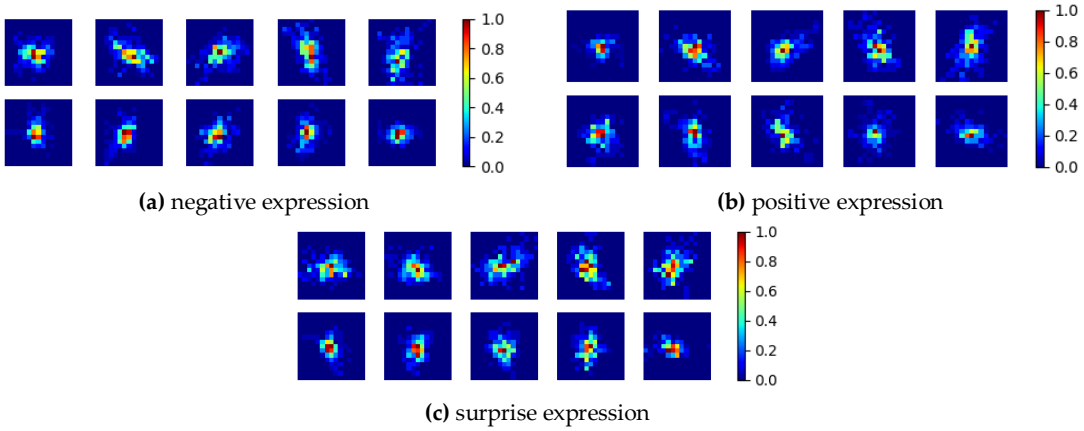

Figura 6. Deplasamentul relativ al centrelor, vizualizat ca un *color map,* pentru fiecare din cele 10 celule analizate. **(a).** micro-expresie de tip negativ. **(b).** micro-expresie de tip pozitiv. **(c).** micro-expresie de tip surpriză.

În faza de testare, pentru a determina tipul unei noi micro-expresii, vectorul de trăsături X este extras din secvență și, pentru fiecare celulă, se calculează probabilitatea ca mișcarea apărută să aparțină claselor micro-expresiilor. Pentru aceasta utilizăm funcția de distribuție de probabilitate normal multi-variată (*multi-variate normal distribution function*):

(7)

În final, probabilitatea ca o secvență să aparțină fiecărei micro-expresii (p_{sur}, p_{neg} sau p_{pos}) se calculează prin înmulțirea probabilităților calculate pentru fiecare celulă. Tipul micro-expresiei este selectat ca fiind maximul p_{sur}, p_{neg} and p_{pos}:

(8)

, unde este probabilitatea ca deplasamentul relativ al centrelor corespunzător celulei *c* să aparțină micro-expresie *e*. Peste probabilitățile inițiale se aplică funcția logaritm pentru a asigura stabilitatea numerică a rezultatului.

Metoda propusă a obținut o rată de detecție de 79.23% și o rată de recunoaștere de 82.59% .

În ultima perioadă, rețelele convoluționale neuronale au început să fie folosite pe scară largă și aproape toate problemele din domeniul viziunii computerizate au fost abordate și s-au obținut rezultate impresionante cu aceste rețele. Am propus un framework original pentru detecția și recunoașterea micro-expresiilor bazat pe rețele neuronale convoluționale. Rețeaua neuronală selectează automat trăsăturile relevante din imaginea de intrare și realizează clasificarea. Arhitectura soluției propuse este prezentată în Figura 7.

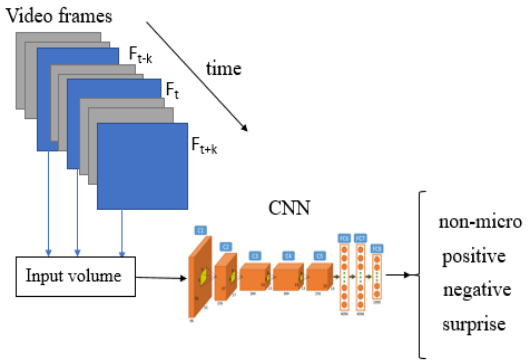

Figura 7: Arhitectura soluției de analiză a micro-expresiilor pe baza rețelelor neuronale convoluționale

Pentru început vom defini conceptele implicate în soluția propusă. Vom folosi o fereastră temporală glisantă pentru a parcurge în mod iterativ fluxul video de intrare. Pe baza vitezei de achiziție a cadrelor (*frame rate*) a dispozitivului de achiziție, calculăm numărul mediu de cadre Δt în care o micro-expresie este vizibilă. Acest parametru a fost ales pe baza duratei fiziologice a unei micro-expresii: 1/15 dintr-o secundă. La fiecare pas, inspectăm cadrul curent *Ft* cu scopul de a determina dacă o micro-expresie a survenit în acest cadru și pentru a o recunoaște dacă este necesar. De asemenea, folosim cadrele *onset Ft - k* și *offset Ft + k* cu scopul de a extrage trăsăturile de mișcare, unde $k = (\Delta t + 1)/2$. Primele și ultimele *k* cadre din fluxul video sunt excluse datorită faptului că în acest caz cadrele *onset* și *offset* vor depăși limitele filmului video.

Fiecare cadru din filmul video de intrare, *Ft*, împreună cu cadrele sale *onset* și *offset* corespunzătoare, sunt introduse ca intrări ale unei rețele neuronale convoluționale (AlexNet), care va clasifica starea la momentul de timp *t*, într-una din următoarele clase: micro (pozitivă,

negativă, surprindere) sau non-micro-expresie. Răspunsurile primare ale rețelei neuronale convoluționale sunt apoi procesate în continuare pentru a se stabili cadrul temporal exact la care a apărut micro-expresia și pentru a filtra răspunsurile greșite de tip *false positives*.

În general se acceptă faptul că micro-expresiile corespund celor șapte emoții universale: surprindere, mânie, frică, tristețe, dezgust, dispreț, fericire. Totuși, întrucât rețelele neuronale convoluționale au nevoie de volume mari de date de antrenament, am ales o taxonomie cu numai trei clase: micro-expresie pozitivă, negativă și de surprindere, pentru partea de recunoaștere a micro-expresiilor.

Soluția propusă atinge o acuratețe globală de 72.2%.

3.3. Proiectarea și implementarea algoritmilor pentru recunoașterea atributelor faciale

Analiza fețelor umane ocupă o poziție privilegiată în diverse domenii pluridisciplinare deoarece fețele umane exprimă diverse informații: atribute demografice (vârstă, rasă, gen, etnie etc.), semnale sociale, emoții, și lista rămâne deschisă.

În cadrul acestei activități ne-am focalizat în principal pe recunoașterea trăsăturilor demografice pe baza imaginilor faciale. Mai precis am propus mai multe metode pentru: (1) recunoașterea rasei și a etniei. (2) recunoașterea genului și (3) recunoașterea culorii pielii.

Recunoașterea automată a grupurilor etnice și rasiale are implicații importante în cadrul unei game largi de discipline, cum ar fi medicina, interacțiunea om-calculator (HCI – *human computer interaction*), biometria, sistemele de supraveghere, visagismul (*visagisme*) etc. De exemplu, medicina bazată pe rase și farmacogenomica orientată spre rase (*race targeted pharmacogenomics*) promovează utilizarea informațiilor cu privire la rase în diagnoza și tratamentul mai multor maladii. Informațiile de biometrie soft (așadar, aici intră și cele cu privire la rasă) pot fi încapsulate în sistemele de supraveghere video în vederea îmbunătățirii acurateței procesului de identificare a persoanelor, reducând drastic numărul de potriviri posibile. De asemenea, informațiile cu privire la rase pot fi folosite în aplicații HCI și în cadrul sistemelor de „publicitate țintită" (*targeted advertising*) pentru a oferi utilizatorilor servicii adecvate din punct de vedere etnic, evitând astfel situațiile în care utilizatorii s-ar putea simți jigniți datorită unor tabuuri culturale. Desigur, se pot avea în vedere și multe alte aplicații.

Sistemul propus pentru recunoașterea rasei și a etniei este ilustrat în Figura 8.

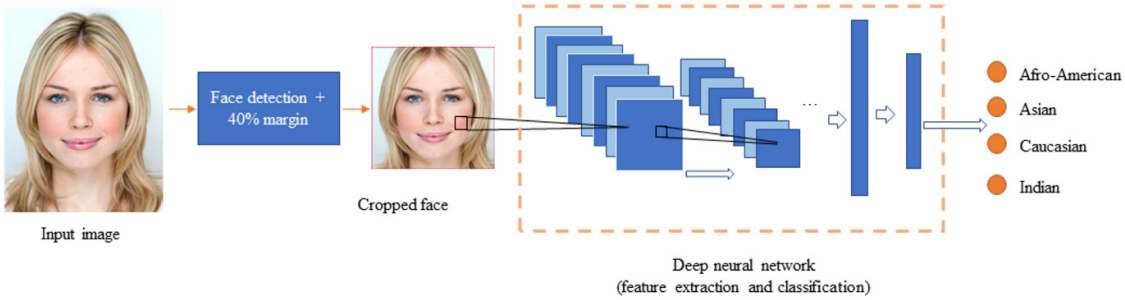

Figura 8. Arhitectura sistemului pentru detecția raselor umane

Această sistem pune în evidență următoarele contribuții:

- Structurarea unui set de date „*in the wild*" de mari dimensiuni conținând chipuri umane adnotate cu informații referitoare la rasă și la apartenența etnică. După

cunoștințele noastre, noi am structurat cea mai mare bază de date de chipuri umane disponibilă (cuprinzând peste 200.000 de imagini), adnotată cu informații referitoare la rasă și la apartenența etnică. Etichetarea apartenenței etnice este realizată numai pentru subiecți asiatici care sunt separați mai departe în chinezi, japonezi și coreeni.

- Antrenarea și compararea a patru rețele neuronale convoluționale „*state of the art*" (CNN – *convolutional neural networks*) pe un *use-case* specific: cel al clasificării rasiale. Taxonomia pe care o propunem conține patru etichete rasiale: asiatic, negru, caucazian și indian. Cea mai bună performanță a fost obținută de Inception Resnet-v2 (96.36%), pe când Alexnet obține cea mai slabă performanță (94.53%), diferența dintre ele fiind de numai 1.83%.
- Apoi, utilizând tehnica de *transfer learning*, rețelele antrenate pentru clasificare rasială sunt acordate fin pentru clasificare etnică între subiecții de origine asiatică. Mai specific, ultimele straturi ale rețelelor sunt reantrenate astfel încât rețelele să poată distinge între subiecți chinezi, japonezi și coreeni.
- În final se folosesc multiple tehnici de vizualizare pentru a „vedea" ce anume au învățat rețelele și pentru a realiza o discuție comparativă cu privire la modul în care ființele umane și rețelele convoluționale percep rasa.

Pentru a exprima numeric sensibilitatea rețelei față de anumite trăsături faciale, au fost aplicate mai multe transformări (ocluzii, efect de ceață, accentuări coloristice – Figura 9) asupra imaginilor de test pentru a degrada cele mai proeminente trăsături faciale. Imaginile modificate sunt apoi introduse în CNN și rezultatele sunt re-examinate. Regiunea perioculară pare a avea cel mai mare impact asupra performanței clasificării: acuratețea globală scade cu 5.99% atunci când ochii sunt încețoșați și cu 15.62% când ochii sunt complet acoperiți. Zona gurii și a nasului par a avea o importanță mai mică. Acuratețea obținută în cazul clasificării subiecților asiatici și caucazieni este afectată mai ales de alterarea zonei ochilor, deoarece cantusul (unghiul ochiului, unghiul palpebral, unghiul fantei palpebrale) este esențial pentru diferențierea subiecților caucazieni de cei aparținând celorlalte rase.

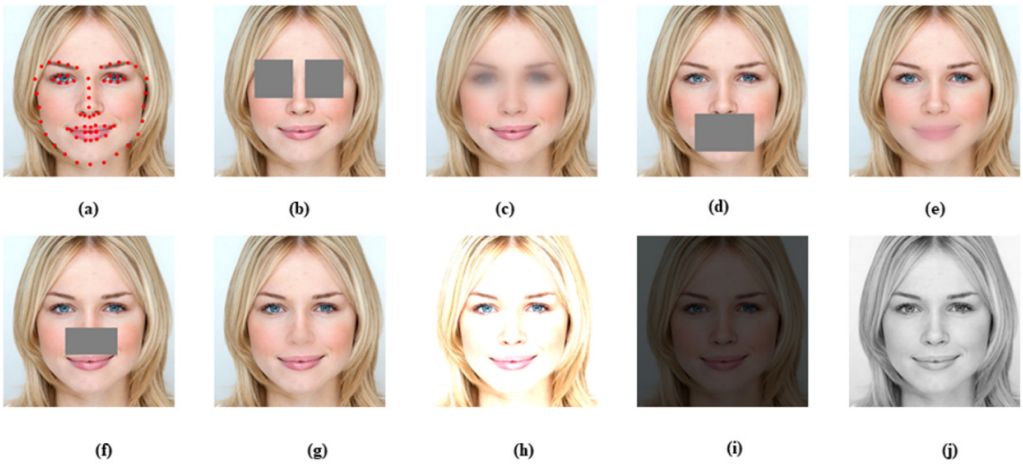

Figura 9. Transformările aplicate pe imagini

Aceste observații confirmă modul în care ființele umane percep fețele [12, 13]: studiile au revelat importanța regiunii perioculare, urmată de zona gurii și apoi de cea a nasului pentru perceperea și recunoașterea chipului uman. Alte studii sugerează că sprâncenele ar putea fi chiar mai importante decât ochii [14], datorită rolului lor în

comunicarea non-verbală și datorită faptului că ele constituie niște trăsături faciale de mari dimensiuni și care apar frecvent în imagini.

Relația geometrică [14] dintre zonele faciale este cel puțin la fel de importantă ca aparența fiecărei trăsături faciale. Deși uneori trăsăturile singure sunt suficiente pentru recunoașterea facială, „relația geometrică dintre fiecare trăsătură și restul feței se poate dovedi mai importantă decât gradul de participare la stabilirea diagnosticului al trăsăturii respective". Aceasta ar putea fi o explicație a faptului că performanța procesului de detecție nu scade drastic atunci când se maschează în mod independent diverse părți ale feței.

Informațiile cromatice nu par a avea un impact prea mare: când imaginile sunt convertite în tonuri de gri (*grayscale*), performanța globală scade cu 1.9%.

Experimentele pe care le-am realizat arată că sistemul propus este robust, invariant la iluminare și demonstrează acuratețea soluției.

Pentru detecția genului din imagini faciale am folosit o metodologie asemănătoare: în primul rând s-a asamblat o bază de date cu imagini faciale de pe Internet (60000 de imagini faciale) ; fețele au fost detectate în fiecare imagine și pentru fiecare față s-a stabilit genul persoanei. Apoi s-au antrenat și evaluat mai multe rețele convoluționale neuronale pe imaginile colectate de pe internet. Cele mai bune rezultate au fost obținute de rețeaua VGG-19 cu o acuratețe de 97.7%.

În ceea ce privește determinarea culorii pielii am propus o abordare complet automată care nu necesită nici o calibrare prealabilă a camerei de luat vederi, pentru a clasifica culoarea pielii în trei clase: închisă, medie și deschisă. Metoda a fost dezvoltată pentru un sistem optic de visagisme (*visagisme*) în care culoarea pielii este analizată pentru a-i sugera utilizatorului cei mai potriviți ochelari care sunt în armonie cu fața sa. Am propus două metode de clasificare a culorii pielii. Prima metodă folosește calculul histogramelor de culoare în diferite spații de culoare, după care le concatenează într-un vector de trăsături. Dimensionalitatea vectorului de trăsături este redusă cu ajutorul *Principal Component Analysis*, iar rezultatul este transmis ca intrare unui clasificator de tip *Support Vector Machine* (SVM) pentru a determina nuanța pielii. Cea de-a doua metodă folosește rețele neuronale convoluționale pentru a clasifica nuanța de culoare a pielii; în acest mod, trăsăturile cromatice relevante sunt extrase în mod automat de către rețea. Algoritmii au fost antrenați și testați pe imagini culese din domeniul public, din seturi de date astfel disponibile; metoda de clasificare bazată pe SVM obține o acuratețe de 89.53%, iar rețeaua neuronală convoluțională obține o acuratețe de 91.29%.

Se dorește ca abordarea propusă să fie integrată într-un sistem de analiză a atributelor faciale folosit pentru încercări virtuale de ochelari.

În prima fază se captează o imagine facială a subiectului, după care sistemul (mai precis modulul de *Extragere a atributelor faciale*) determină în mod automat culoarea pielii (precum și alte atribute demografice: gen, vârstă, culoarea ochilor etc.). Pe baza acestor atribute, modulul de *Selecție a Cadrelor* realizează o interogare a bazei de date de ochelari 3D și selectează accesoriile care sunt în armonie cu fața utilizatorului. Fiecare pereche de ochelari 3D a fost adnotată în prealabil de către un specialist în visagism (*visagisme*), care atribuie câte un scor fiecărui atribut facial; utilizatorului nu îi sunt afișați sau prezentați decât ochelarii cu cele mai mari scoruri. În mod tipic, setul de date cuprinzând imagini de ochelari 3D conține mai multe mii de modele de ochelari.

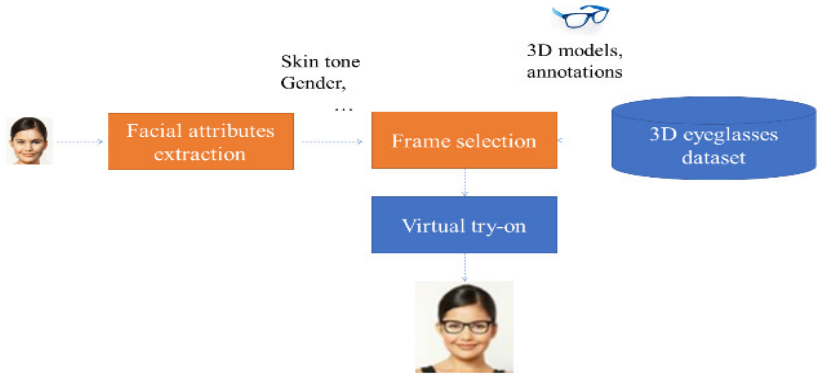

Figura 10. Sistemul de visagism pentru analiza trăsăturilor faciale

Desigur, se pot avea în vedere și alte aplicații: datele extrase de către modulul de *Extragere a atributelor faciale* pot fi folosite, de pildă, pentru a sugera culoarea cea mai adecvată a rujului, fondului de ten sau a vopselei de păr.

3.4. Realizarea aplicației demonstrator

Am conceput un sistem fizic pentru obținerea și captarea micro-expresiilor: am utilizat o cameră video de mare viteză Ximea [15] conectată la un calculator PC; PC-ul afișează clipuri video cu o puternică încărcătură emoțională pentru utilizator și înregistrează reacția acestuia la acești stimuli folosind camera de luat vederi cu înaltă rezoluție temporală. Sistemul folosește o cameră de luat vederi de mare viteză Ximea MQ022CG-CM USB3.0 echipată cu o lentilă Fujinon de 9mm (care asigură un câmp vizual orizontal de 64° și un câmp vizual vertical de 37°), capabilă să capteze cadre video de 2048 × 1088 pixeli la o frecvență maximă de 170 de cadre pe secundă. Camera de luat vederi poate fi configurată pentru a folosi achiziție de imagini focalizată pe regiunea de interes (ROI - Region Of Interest) și astfel permite obținerea cadrelor cu frecvențe și mai mari.

Scenariul de utilizare de bază al sistemului propus este următorul (Figura 11): utilizatorul este rugat să se așeze în dreptul camerei de luat vederi și urmează ca le să fie supus unor stimuli susceptibili să declanșeze răspunsuri emoționale (de exemplu, utilizatorul este rugat să vizioneze anumite clipuri video special selecționate).

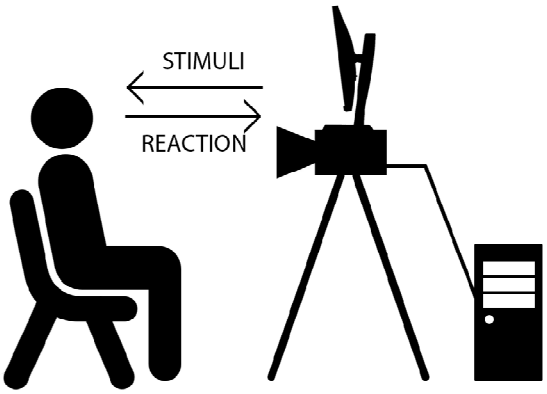

Figura 11. Procesul de achiziție rapidă și de analiză de imagini

Sistemul începe să achiziționeze imagini la rezoluția maximă a cadrului full (2048 x 1088 pixeli) folosind o frecvență de achiziție de 30 de cadre pe secundă. Calculatorul gazdă care realizează procesarea imaginilor detectează automat fața utilizatorului folosind o bibliotecă disponibilă public de detecție a fețelor. După detectarea feței, se stabilește în mod automat regiunea de interes ROI astfel încât aceasta va include fața detectată și o zonă semnificativă de siguranță în jurul acestei fețe (lățimea și înălțimea ROI sunt cu 75% mai mari decât dimensiunile feței detectate). Camera video este configurată pentru a utiliza un algoritm de achiziții de imagini bazat pe ROI folosind ROI detectată, ceea ce reduce în mod semnificativ volumul de date transferate prin USB și stocate în memoria calculatorului, permițând astfel capturi video la viteză foarte mare (peste 110 cadre pe secundă). Rata de achiziție reală depinde nu numai de volumul de date transferate, ci și de timpul de expunere, care nu este influențat de dimensiunea ROI. Diagrama de flux a acestui proces este prezentată în Figura 12.

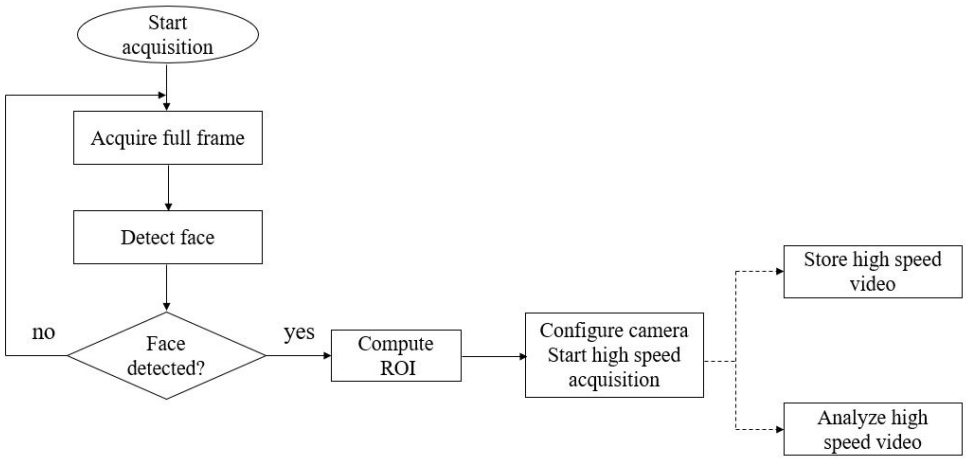

Figure 12. Procesul de achiziție și de analiză video de înaltă viteză

Sistemul poate fi utilizat fie pentru a capta și a stoca secvențele video de înaltă viteză, fie pentru a analiza online (pentru a detecta și a recunoaște expresii și micro-expresii) *stream*-ul video.

Pentru a salva cadrele video de înaltă viteză, am folosit binecunoscutul mecanism de sincronizare producător-consumator (Figura 13). *Thread*-ul producător citește cadrele video de înaltă viteză de la camera de luat vederi Ximea [15] și le salvează în coada de date partajată, în timp ce *thread*-ul consumator citește cadrele din coadă și le salvează pe dispozitivul de stocare fizic. *Thread*-ul consumator are nevoie să interacționeze în mod constant cu sistemul de fișiere, ceea ce este un proces consumator de timp, în timp ce *thread*-ul producător citește cadrele de la camera de luat vederi cu o rezoluție temporală ridicată. Pentru a evita erorile de tipul *out-of-memory*, cadrele și etichetele lor temporale (*timestamps*) corespunzătoare sunt salvate pe hard disc în rafale de *nf* = 100 cadre, direct în format binar folosind apeluri de sistem de nivel coborât.

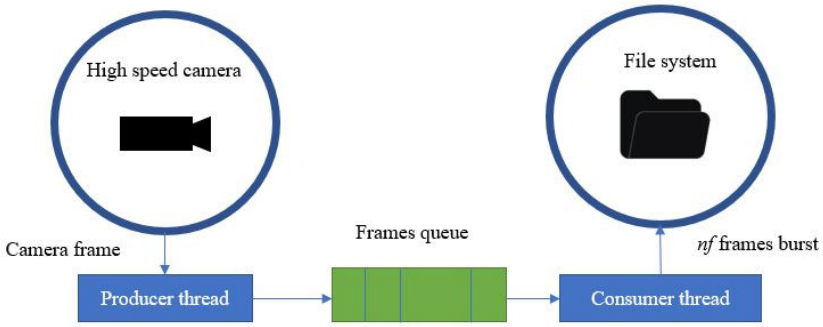

Figura 13. Procesul de captură și stocare video de înaltă viteză

În Tabelul 1 raportăm performanța sistemului fizic realizat în diferite condiții de utilizare; toate experimentele au fost realizate în condiții naturale de iluminare. (mediu interior: Experimentele 1-4 și mediu extern: Experimentele 5-12).

Tabelul 1. Frame-rate-ul obținut în diferite condiții de iluminare

Experiment	Timp de expunere (milisecunde)	FPS	Distanța până la subiect (cm)	Intensitatea luminii (Lux)	Dimensiunea imaginii
1	7	118	80	770	654x654
2	7	138.5	100	770	458x458
3	7	138	120	770	386x386
4	7	139	200	770	246x246
5	5	179	80	1100	497x497
6	5	198	100	1100	404x404
7	5	198	120	1100	360x360
8	5	199	200	1100	239x239
9	4	189	80	1300	542x542
10	4	243	100	1300	423x423
11	4	249	120	1300	358x358
12	4	249	200	1300	255x255

Sistemul propus este capabil să captureze imagini faciale la un frame-rate mai mare de 118 cadre pe secundă în medii interioare și la un frame-rate mai mare de 200 cadre pe secundă în medii exterioare. În primul caz, timpul de expunere trebuie să fie mai lung astfel că apare o scădere a rezoluției temporale. Timpul de expunere a fost determinat euristic, prin multiple încercări, astfel încât imaginea facială să aibă o iluminare optimă.

Crearea unei baze de date cu micro-expresii este un proces îndelungat care necesită cunoștințe de specialitate (psihologie comportamentală, *Facial Action Coding System*); de aceea, sistemul de analiză al micro-expresiilor a fost validat pe baze de date publice deja adnotate. În plus, am evaluat sistemul si pe imagini capturate în laboratorul de cercetare pentru a determina robustețea acestuia față de *false positives*.

3.5. Calibrarea camerelor folosind rețele neuronale convoluționale

Pentru auto-calibrarea sistemului am realizat un sistem nou de detectie a punctului de fuga (“VP”) folosind retele neuronale. Metodele traditionale bazate pe procesarea imaginilor au fost folosite pentru a genera o baza de date, care a fost corectata si ajustata manual unde a fost necesar. Datele sunt apoi impartite in 90% date de antrenare si 10% date de testare. Prin augmentarea datelor am dublat dimensiunea initiala prin simpla oglindire orizontala a imaginilor si ajustarea punctelor de fuga. In cele din urma s-au obtinut 4830 imagini de antrenare, care au fost impartite din nou in 90% date antrenare si 10% date de validare folosite in momentul antrenarii retelei la minimizarea erorii de predictie.

Sistemul de detectie a punctelor de fuga este ilustrat in figura 14:

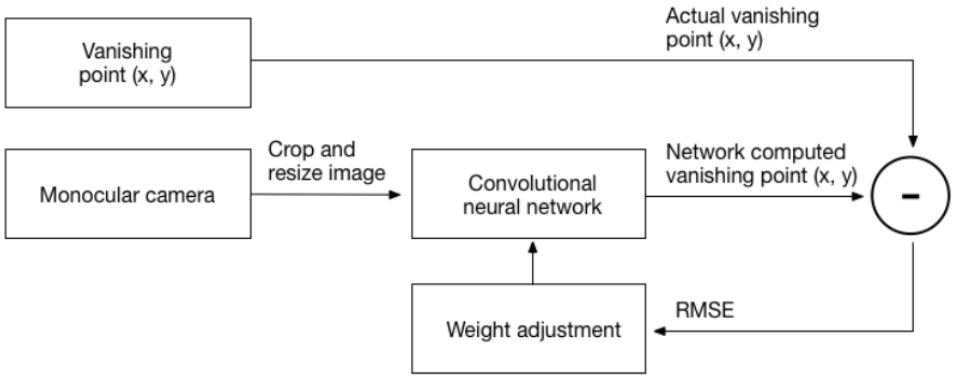

Figura 14. Sistemul de detectie bazat pe retele neuronale.

Modelul retelei neuronale convolutionale este prezentat in figura 15. Unii pasi traditionali cum ar fi: pre-procesarea imaginii, selectia trasaturilor relevante si extragerea lor, iar apoi clasificarea acestor trasaturi sunt eliminate sau inlocuite complet de reletele neuronale convolutionale.

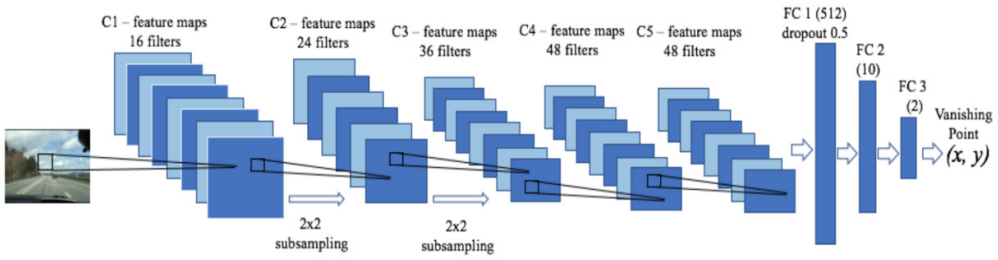

Figura 15. Modelul retelei neuronale convolutionale.

Datele de intrare sunt normalizate si apoi trecute prin 5 “layere” convolutionale (C1-C5) cu numar diferit de filtre si avand kernel de dimensiune variabila. Ultimele “layere” ale retelei sunt reprezentate de FC1-FC3 care sunt denumite “fully-connected layers”, echivalentul unor retele neuronale clasice unde in functie de neuronii de intrare se activea o

anumita iesire. De altfel, ultimul FC3 cu cele 2 valori ale neuronilor reprezinta valorile coordonatelor x si y ale punctului de fuga prezis.

Reteaua neuronala a fost antrenata avand prin minimizarea erorii "RMSE" ("root mean squared error") care in cazul de fata reprezinta distanta in pixeli dintre punctul prezis de retea si punctul de fuga real din baza de date.

Tabelul 2. Evaluarea predictiei de punct de disparitie folosind "RMSE".

Metrica eroare	Scenariu 1	Scenariu 2	Scenariu 3
RMSE	5.19591	8.47350	4.36344

Rezultatele predictiei folosind reteaua neuronala sunt ilustrate in tabelul 2 unde s-a folosit metrica "RMSE" pentru eroare in 3 scenarii diferite.

4. Prezentarea sintetică a celor mai importante contribuții ale proiectului

În această secțiune se vor prezenta contribuțiile principale obținute în urma derulării acestui proiect de cercetare.

4.1. Realizarea unui sistemul sensorial multifocal pentru analiza reacțiilor utilizatorului

Sistemul multifocal pentru achizitia imaginilor in vederea recunoasterii si urmaririi trasaturilor faciale si corporale (Figura 16) utilizează doua camere video montate si calibrate in configuratie de stereoviziune. A treia camera este montata intre cele doua camere stereo si este de viteza mare, dar cu un camp vizual redus. Aceasta camera este montată pe un mecanism unic de tip pan tilt. Mecanismul va folosi doua motoare de tip stepper pentru a controla miscarea de inclinatie si de rotatie a camerei. Folosirea motoarelor stepper va aduce un plus de precizie sistemului de urmarire. Controlul acestor motoare va fi dat de microcontroller-ul Arduino si un controller de motoare conectat la Arduino. Camera se va

monta pe un suport ce urmeaza a fi realizat de echipa de cercetare.

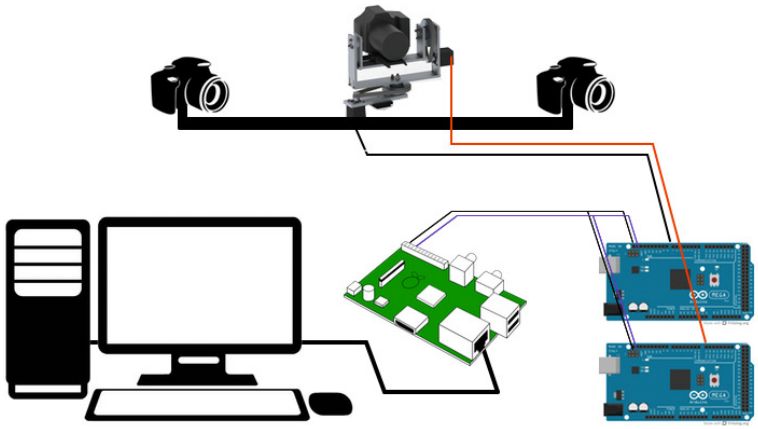

Figura 16. Vedere de ansamblu asupra sistemului multifocal

Pentru a sustine greutatea unei camere video am ales folosirea de motoare pas cu pas, care au cuplu (moment) mare. Controlul motoarelor va fi realizat prin intermediul unui driver de motoare: "Arduino Motor Shield" (https://www.arduino.cc/en/Main/ArduinoMotorShieldR3). Avand un motor pas cu pas de 200 de pasi vom putea controla miscarea de rotatie a acestuia in incremente de 1.8 grade pe o miscare circulara completa de 360 de grade.

Sistemul începe să achiziționeze imagini la rezoluția maximă a cadrului full (2048 x 1088 pixeli) folosind o frecvență de achiziție de 30 de cadre pe secundă. Calculatorul gazdă care realizează procesarea imaginilor detectează automat fața utilizatorului folosind o bibliotecă disponibilă public de detecție a fețelor. După detectarea feței, se stabilește în mod automat regiunea de interes ROI astfel încât aceasta va include fața detectată și o zonă semnificativă de siguranță în jurul acestei fețe (lățimea și înălțimea ROI sunt cu 75% mai mari decât dimensiunile feței detectate). Camera video este configurată pentru a utiliza un algoritm de achiziții de imagini bazat pe ROI folosind ROI detectată, ceea ce reduce în mod semnificativ volumul de date transferate prin USB și stocate în memoria calculatorului, permițând astfel capturi video la viteză foarte mare (peste 110 cadre pe secundă). Rata de achiziție reală depinde nu numai de volumul de date transferate, ci și de timpul de expunere, care nu este influențat de dimensiunea ROI.

4.2. Dezvoltarea, implementarea și validarea unor algoritmi originali pentru segmentarea și urmărirea ochilor

În cadrul acestui modul am proiectat și implementat mai multe modele originale de urmărire și segmentare a ochilor în timp real. Contribuțiile principale sunt :

- Utilizarea unui detector de simetrii circulare: *Fast Radial Symmetry Transfrom* pentru detecția irisului în imagini faciale.

- Propunerea unei metode simple și robuste pentru detecția razei irisului: s-a utilizat derivata Sobel a imaginii pe o regiune în jurul ochilor pentru a accentua tranziția puternică dintre zona irisului și a sclerei. Pentru a elimina zgomotele din această regiune, pe imagine se aplică un filtru Gaussian și apoi primele *k%* cele mai mici valori din derivata Sobel sunt ignorate. Pentru fiecare rază candidat din intervalul $[r_{min}, r_{max}]$, se calculează o proiecție radială prin însumarea valorilor gradientului care sunt la o distanță r de centrul irisului și cu o deschidere angulară între $[\theta_{min}, \theta_{max}]$. Această proiecție a imaginii atinge valoarea maximă la granița dintre regiunea sclerei și a irisului.
- Propunerea unui descriptor original bazat pe două parabole (una pentru pleoapa de sus și una pentru pleoapa de jos) pentru descrierea formei exterioare a ochiului (Figura 17).

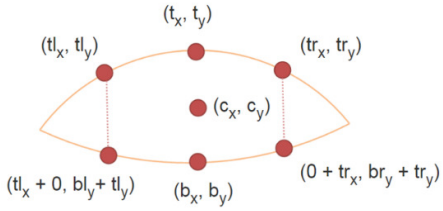

Figura 17. Descriptorul utilizat pentru descrierea formei exterioare a ochilor

- Propunerea unui algoritm de segmentare a formei exterioare a ochilor folosind informații de culoare. Am antrenat un clasificator de tipul Support Vector Machine (*SVM*) pe regiuni (*patch*) din zona ochilor pe imagini dintr-o bază de date publică. Trăsăturile utilizate pentru învățare sunt: canalul de *Hue* din spațiul de culoare HSV și canalele O1 și O2 din spațiul de culoare RGB opus (*RGB opponent*).
- Pentru potrivirea unei forme ipotetice la imagine s-a definit o metodologie originală bazată pe metode Monte Carlo. Se analizează pixelii pozitivi (p_+), care reprezintă pixelii din regiunea sclerei și pixelii negativi (p_-) care reprezintă pixelii din zona pielii și a genelor. Procesul de potrivire (doar pentru pleoapa de sus) este ilustrat în figura 4. Scorul de potrivire al unei ipoteze este definit ca:

---

unde α și β ($\alpha + \beta = 1$) sunt două ponderi care determină influența pixelilor pozitivi și respective negative.

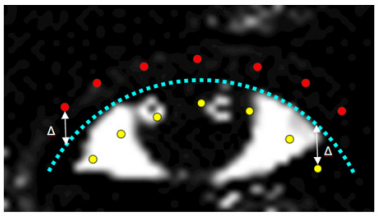

Figura 18. Procesul de potrivire al unei ipoteze: linia punctată reprezintă conturul pleoapei de sus al unei ipoteze, iar pixelii pozitivi sunt marcați cu galben și pixelii negativi cu roșu.

- Extinderea algoritmului descris mai sus pentru a fi invariabil la condițiile de iluminare. Pentru procesul de potrivire s-au utilizat și informații legate de colțuri (în jurul colțurilor ochilor ar trebui ca un detector general de colțuri să aibă un

răspuns puternic) și de derivatele imaginii. De asemenea potrivirea unei forme ipotetice s-a făcut cu filtre de particule.

- Dezvoltarea unei metode originale care să permită urmărirea ochilor în cadrele video. Soluția propusă folosește o abordare de la grosier la fin (*coarse-to-fine*) pentru a detecta și urmări multiple trăsături ale ochilor: centrul irisului, conturul ochiului și clipirile. Soluția utilizează trei filtre de particule în paralel pentru a urmări ochii: primul filtru de particule este folosit pentru a determina pozițiile aproximative ale irisurilor. Alte două filtre de particule sunt utilizate pentru a determina și a urmări conturul fiecărui ochi pe baza estimării obținute de la primul filtru de particule.

4.3. Dezvoltarea, implementarea și validarea unor algoritmi originali pentru detecția și recunoașterea micro-expresiilor din secvențe video de mare viteză

Micro-expresiile sunt expresii faciale scurte, cu o durată cuprinsă între 1/12 și 1/25 secunde, și ele sunt surse sigure de detecție a înșelăciunii. Există două probleme majore care trebuie să fie adresate în detecția micro-expresiilor. În primul rând, acestea sunt mișcări involuntare, deci este dificil să se obțină datele de test. La ora actuală sunt disponibile mai multe baze de date cu micro-expresii. A doua problemă o constituie faptul că acestea sunt vizibile doar un număr redus de cadre, deci necesită algoritmi preciși de urmărire și detecție a mișcării.

Am adus următoarele contribuții originale metodelor de urmărire a microexpresiilor:

- Proiectarea și implementarea unui framework original pentru detecția și recunoașterea micro-expresiilor bazat pe rețele neuronale convoluționale. Rețeaua neuronală selectează automat trăsăturile relevante din imaginea de intrare și realizează clasificarea. Rețeaua primește ca intrare două imagini diferență: fiecare cadru din filmul video de intrare, *Ft*, împreună cu cadrele sale *onset* și *offset* corespunzătoare, sunt introduse ca intrări ale unei rețele neuronale convoluționale.
- Proiectarea și implementarea unui algoritm generic pentru post-procesarea răspunsului clasificării per cadru a unui sistem automat de detecție al micro-expresiilor (detecția clasifică frame-urile video în cadru care conține o micro-expresie sau cadru non-micro expresie). Experimentele efectuate arată ca algoritmul propus îmbunătățește semnificativ procesul de detecție și *false-positive-urile* sunt în general filtrate.
- Propunerea unei noi metode de selecție a regiunilor relevante de pe față implicate în expresia emoțiilor. Pentru aceasta ne-am bazat pe studiul mușchilor faciali implicați în producerea expresiilor faciale.
- Propunerea unui nou descriptor imagine (*Movement Magnitude* image) pentru detecția mișcărilor subtile apărute în cadrul micro-expresiilor. Întrucât scopul modulului de detecție este de a găsi cadrele apex, considerăm diferența imagistică absolută dintre cadrul curent t (un cadru apex potențial) și cadrul anterior de la o distanță $\tau/2$ (un cadru onset potențial). Cu toate acestea, întrucât mișcările faciale care survin în timpul unei micro-expresii au o intensitate foarte scăzută, am introdus și un factor de normalizare pentru a distinge mișcarea de tip ME de zgomotul cauzat de condițiile de iluminare sau de dispozitivele de captare. Cadrul $t - \varepsilon$ ($\varepsilon = 3$ în experimentele noastre) este folosit ca factor de normalizare. Deoarece secvențele video sunt capturate cu camere de înaltă viteză, nici o mișcare facială nu ar trebui să se producă în 0.03 s (valoare calculată pentru o rezoluție temporală de 100 cadre pe secundă).

În final imaginea *movement magnitude* este definită ca:

- Proiectarea și implementarea unui algoritm rapid și simplu (bazat doar pe imaginea *Movement Magnitude*) pentru detecția microexpresiilor. Metoda se bazează pe algoritmi de învățare *ensemble. (ensemble learning methods)*
- Propunerea unui nou descriptor de mișcare pentru recunoașterea tipului de micro-expresie. Descriptor se bazează tot pe imaginea *movement magnitude* (MM). Pentru fiecare celulă se calculează poziția ponderată centroidului pe baza intensității fiecărui pixel din imaginea movement magnitude (MM) și pe regiunile selectate de pe față:

, unde reporezintă suma pixelilor din imaginea *MM* dintr-o celulă, *MM(x, y)* este valoarea pixelului din imaginea MM de la poziția (x, y) iar c_s, r_s, c_M, r_M definesc regiunea unei celule (*bounding rectangle*).

- Propunerea unei metode de recunoaștere a tipului micro-expresie care nu necesită învățare și care utilizează poziția ponderată a centroidului definită mai sus.

4.4. Dezvoltarea, implementarea și validarea unor algoritmi originali pentru extragerea trăsăturilor faciale

Am propus mai mulți algoritmi pentru extragerea atributelor faciale: extragerea culorii pielii, extragerea rasei și a etniei și extragerea genului. Aceste metode au o aplicație practică: ele sunt integrate într-un sistem de analiză a atributelor faciale folosit pentru încercări virtuale de ochelari.

În prima fază se captează o imagine facială a subiectului, după care sistemul (mai precis modulul de *Extragere a atributelor faciale*) determină în mod automat culoarea pielii (precum și alte atribute demografice: gen, vârstă, culoarea ochilor etc.). Pe baza acestor atribute, modulul de *Selecție a Cadrelor* realizează o interogare a bazei de date de ochelari 3D și selectează accesoriile care sunt în armonie cu fața utilizatorului. Fiecare pereche de ochelari 3D a fost adnotată în prealabil de către un specialist în visagism (*visagisme*), care atribuie câte un scor fiecărui atribut facial; utilizatorului nu îi sunt afișați sau prezentați decât ochelarii cu cele mai mari scoruri. În mod tipic, setul de date cuprinzând imagini de ochelari 3D conține mai multe mii de modele de ochelari.

Desigur, se pot avea în vedere și alte aplicații: datele extrase de către modulul de *Extragere a atributelor faciale* pot fi folosite, de pildă, pentru a sugera culoarea cea mai adecvată a rujului, fondului de ten sau a vopselei de păr.

Contribuțiile principale aduse în cadrul acestui modul sunt:

- Realizarea și publicarea unei baze de date de mari dimensiuni (**60000** de imagini faciale) adnotate cu genul persoanei
- Antrenarea și evaluarea mai multor rețele neuronale convoluționale pentru problema determinării genului din imagini faciale

- Realizarea și publicarea unei baze de date de mari dimensiuni (**200000** de imagini faciale) adnotate cu rasa persoanei și cu etnia (doar pentru rasa Asiatică, se propune o taxonomie etnică cu următoarele clase: chinez, corean și japonez)
- Realizarea unui studiu amplu despre modul în care oamenii percep rasele umane
- Antrenarea și compararea a patru rețele neuronale convoluționale „*state of the art*" (CNN – *convolutional neural networks*) pe un *use-case* specific: cel al clasificării rasiale. Taxonomia pe care o propunem conține patru etichete rasiale: asiatic, negru, caucazian și indian. Cea mai bună performanță a fost obținută de Inception Resnet-v2 (96.36%), pe când Alexnet obține cea mai slabă performanță (94.53%), diferența dintre ele fiind de numai 1.83%.
- Dezvoltarea și implementarea mai multor tehnici de vizualizare pentru a „vedea" ce anume au învățat rețelele și pentru a realiza o discuție comparativă cu privire la modul în care ființele umane și rețelele convoluționale percep rasa.
- Compararea, pe baza tehnicilor de vizualizare implementate, a modului în care sistemul bazat pe rețele convoluționale neuronale percepe rasa cu modul în care oamenii percep rasele umane.
- Dezvoltarea unei metode originale care nu necesită calibrare color a camerelor pentru detecția culorii pielii folosind support vector machines, metode de reducere a dimensionalității datelor și histograme color.

5. Lista publicațiilor

Pentru a face publice cele mai importante realizări tehnice pe care le-am obținut am elaborat și am trimis spre publicare mai multe articole în reviste și conferințe de specialitate. Concret s-a publicat un articol într-un jurnal ISI (MDPI Sensors, IF - 2.67) și 9 articole la conferințe internaționale de specialitate. De asemenea, două articole sunt încă în recenzie la două jurnale ISI : PlosOne (IF – 2.806) și MDPI Sensors (IF - 2.67). Lista publicațiilor este prezentată mai jos.

1. Borza, D., Darabant, A. S., & Danescu, R. (2016). Real-Time Detection and Measurement of Eye Features from Color Images. Sensors, 16(7), 1105. [**ISI**]
2. M. P. Muresan, S. Nedevschi & R. Danescu. (2016) "Patch warping and local constraints for improved block matching stereo correspondence," *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2016, pp. 321-327, doi: 10.1109/ICCP.2016.7737167
3. Borza, D. & Danescu, R. (2016). Eye Shape and Corners Detection in Periocular Images Using Particle Filters. In Procedings of the *12th International Conference on Signal Image Technology & Internet Based Systems (SITIS)*, 28 Nov – 1 Dec, Naples, Italy.
4. Borza, D., Darabant, A. S., & Danescu, R. "Fast Eye Tracking and Feature Measurement Using a Multi-Stage Particle Filter". In Procedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), 27 Febr – 1 March 2017.
5. M. P. Muresan, S. Nedevschi & R. Danescu. "A Multi Patch Warping Approach For Improved Stereo Block Matching". In *Procedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 27 Febr – 1 March 2017.

6. Diana Borza, Razvan Itu, Radu Danescu, „Micro expression detection and recognition from high speed cameras using convolutional neural networks", *VISAPP 2018 : International Conference on Computer Vision Theory and Applications*, 27-29.01.2018, **acceptat**.
7. Diana Borza, Adrian Darabant, Radu Danescu, Automatic skin tone extraction for visagsime applications, *VISAPP 2018 : International Conference on Computer Vision Theory and Applications*, 27-29.01.2018, **acceptat**.
8. Sergiu Cosmin Nistor,Alexandra-Cristina Marina, Adrian Sergiu Darabant, Diana Borza, „Automatic gender recognition for "in the wild" facial images using convolutional neural networks", *IEEE Intelligent Computer Communication and Processing (ICCP) 2017*, 7-9.09.2017, pp. 1-5.
9. Diana Borza, Razvan Itu, Radu Danescu, Real-Time Micro-Expression Detection From High Speed Cameras, *IEEE Intelligent Computer Communication and Processing (ICCP) 2017*, 7-9.09.2017, pp. 1-5.
10. Razvan Itu, Diana Borza, Radu Danescu, „Automatic extrinsic camera parameters calibration using Convolutional Neural Networks", *IEEE Intelligent Computer Communication and Processing (ICCP) 2017*, 7-9.09.2017, pp. 1-6.

Articole trimise spre publicare (în recenzie):

1. Diana Borza, Adrian Darabant, Radu Danescu, „Unconstrained race and ethnicity recognition using convolutional neural networks", PLOS One, **IF 2.806, în review**.
2. Diana Borza, Radu Danescu, Razvan Itu, Adrian Darabant, „Micro-expression analysis in high speed video sequences", Sensors, **IF 2.677, în review**.

Bibliografie

1. Cox, M.; Nuevo-Chiquero, J.; Saragih, J.; Lucey, S. CSIRO face analysis SDK. Brisbane, Australia 2013.
2. Saragih, J.M.; Lucey, S.; Cohn, J.F. Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 2011, 91, 200–215.
3. Kazemi, Vahid; Sullivan, Josephine. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014. p. 1867-1874.
4. Kinect face tracking, available online: https://msdn.microsoft.com/en-us/library/jj130970.aspx (Accessed: 19.11.2017)
5. Intel Real Sense Technology, available online: https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html (Accessed: 19.11.2017)
6. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–6.
7. Yan,W.J.; Wu,Q.; Liu,Y.J.; Wang,S.J.; Fu,X. CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–7.
8. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* **2014**, 9, e86041.
9. Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. **2009.**
10. Freund Y.; Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. In European conference on computational learning theory, Mar, 1995, pp. 23-37.
11. Zhu J, Zou H, Rosset S, Hastie T. Multi-class adaboost. *Statistics and its Interface*, **2009**, 2, 3, pp: 349-60.
12. Fraser IH, Craig GL, Parker DM. Reaction time measures of feature saliency in schematic faces. Perception. 1990;19(5):661.
13. Sinha P, Balas B, Ostrovsky Y, Russell R. Face recognition by humans: Nineteen results all computer vision researchers should know about. Proceedings of the IEEE. 2006;94(11).
14. Sadr J, Jarudi I, Sinha P. The role of eyebrows in face recognition. Perception. 2003;32(3)
15. Ximea, Ximea Cameras Homepage, available online: https://www.ximea.com/ , Accessed: 20.10.2017.
16. G. Loy and E. Zelinsky. "A Fast Radial Symmetry Transform for Detecting Points of Interest". In Proceedings of the 7th Euproean Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; p. 358.