



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

Raport științific, Etapa 1.

Anul: 2015

Director proiect: Conf. Dr. Ing. Radu Dănescu

Echipa de cercetare:

Radu Dănescu, Florin Oniga, Diana Borza, Razvan Itu, Mircea Muresan

Cuprins

1. Introducere – Rezumatul etapei.....	2
2. Realizarea sistemului senzorial multifocal: Analiza cerințelor și elaborarea specificațiilor pentru sistemul multifocal)	2
3. Dezvoltarea metodologiei de calibrare preliminară pentru sistemul senzorial multifocal: Studiul soluțiilor de calibrare existente, elaborarea specificațiilor pentru sistemul de calibrare	7
4. Evaluarea acurateții algoritmilor de stereoviziune pentru măsurători ale capului și feței.....	8
5. Proiectarea modelelor pentru geometria și poziția capului, și pentru trăsăturile faciale: Studiul modelelor și a tehnicilor de urmărire existente pentru cap și trăsături faciale	14

1. Introducere – Rezumatul etapei

În aceasta etapa au demarat activitatile in vederea realizarii obiectivelor proiectului: realizarea si calibrarea sistemului multifocal, realizarea sistemului de stereoviziune, și modelarea si urmarirea capului si a trasaturilor faciale. Datorita timpului scurt disponibil pentru aceasta etapa, activitatile esentiale planificate initial in faza de propunere au fost impartite in sub-activitati, in acest an fiind realizate sub-activitatile preliminare, urmand ca in etapa urmatoare activitatile sa fie finalizate.

Astfel, in aceasta etapa a fost realizat un studiu in vederea realizarii sistemului senzorial multifocal, si au fost facute primele achizitii, continand componente de mica valoare, urmand ca achizitia componentelor principale sa fie facuta in etapa 2. A fost definita de asemenea metodologia de calibrare, si au fost studiate performantele a multipli algoritmi de stereoviziune. A fost facut de asemenea un studiu detaliat privind modelarea si urmarirea trasaturilor faciale, definind si potentialele contributii originale punctuale care pot fi aduse.

In vederea constituirii echipei independente de cercetare, in aceasta etapa au fost angajati doi cercetatori doctoranzi care nu au mai avut contract de munca cu Universitatea Tehnica, si un cercetator masterand. In etapele urmatoare, in functie de fondurile disponibile, echipa se va extinde.

Conform planului de realizare, activitatea de cercetare desfasurata in aceasta etapa raspunde cerintelor proiectului, si consideram ca obiectivele planificate pentru acest an sunt indeplinite.

2. Realizarea sistemului senzorial multifocal: Analiza cerințelor si elaborarea specificațiilor pentru sistemul multifocal)

Sistemul multifocal pentru achizitia imaginilor in vederea recunoasterii si urmaririi trasaturilor faciale si corporale (fig. 1) va folosi 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 va fi montata 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.

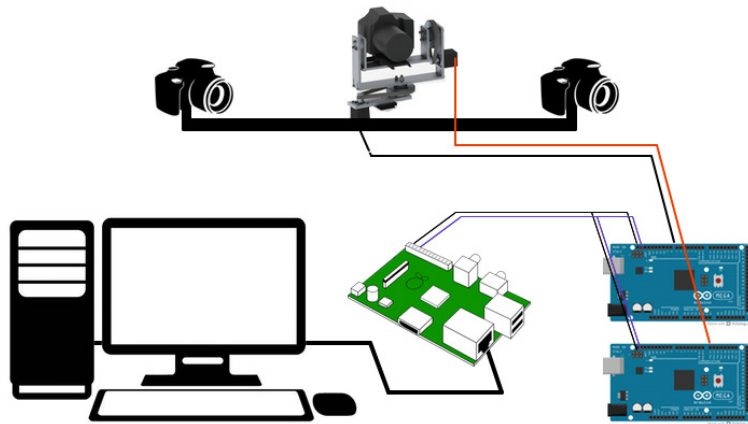


Fig. 1. Vedere de ansamblu asupra sistemului multifocal

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.

Pentru interfața cu driverul motor se folosesc plăci de tip Arduino, care pot fi conectate direct la un sistem de tip PC, sau la un microcalculator de tip pe Raspberry Pi 2 (<https://www.raspberrypi.org/products/raspberry-pi-2-model-b/>), care poate asigura și o procesare intermediară a semnalului și a comenzilor. Raspberry Pi este un computer de dimensiune redusa de tipul "SBC" (Single Board Computer - computer pe o singura placa). Dispune de 4 port-uri USB si de o mufa Ethernet care se poate folosi pentru a comunica cu rețele externe sau alte computere. Datorita posibilitatii usoare de conectare cu alte medii s-a ales folosirea platformei Raspberry Pi, care totodata este si foarte accesibila ca si cost de achizitie.

Pentru execuția algoritmilor de viziune artificială specifici proiectului, există mai multe variante de platforme de calcul. Daca dimensiunea fizica a sistemului nu reprezinta o problema, o alegere buna o poate reprezenta un computer de tip desktop cu procesor modern care are unitate de prelucrare video integrata (GPU), iar daca bugetul permite se poate opta si pentru un sistem cu o placa video dedicata pentru procesare video. O alegere populara o reprezinta si laptop-urile deoarece au avantajul portabilitatii si al mobilitatii si faptul ca ofera un monitor pentru vizualizarea rezultatelor. Daca totusi dimensiunea unui dekstop sau laptop ar fi prea mare, o varianta mai adecvata ar fi alegerea unui sistem de tip embedded. Tabelul 1 prezintă o sinteză a opțiunilor de calcul.

Tabel 1. Soluții de calcul pentru execuția algoritmilor.

Tip/familie placa	Putere calcul	Consum curent	Caracteristici
Microcontroller	< 0.2 GFLOPS	< 0.3 W	Acces firmware, functii low-level, pini GPIO, necesita alte componente pentru

			vizualizare rezultate
SOC mobil sau tableta	1 - 25 GFLOPS	1 - 6 W	Dimensiuni reduse, acces pini GPIO, necesita alte componente pentru vizualizare rezultate
SBC x86, laptop sau netbook	15 - 110 GFLOPS	30 - 100 W	Dimensiuni medii, usor de programat, unele necesita monitor pentru vizualizare
Laptop x86 cu GPU dedicat	240 - 2200 GFLOPS	40 - 110 W	Cel mai performant si usor de programat, dimensiune mare, costuri mari de achizitie
Hardware FPGA	50 - 1000 GFLOPS	1 - 3 W	Foarte performant, dimensiune redusa, dar greu de implementat

Daca se alege varianta microcontroller, atunci consumul de curent este foarte redus, sub 0,3 Watt, dar puterea de procesare nu este asa de mare: pana in 0.2 GFLOPS. Majoritatea microcontrollerelor populare (ex: Arduino, AVR, PIC) sunt prea slabe pentru a procesa imagini dintr-o camera video. Totusi, unele placi bazate pe ARM Cortex-M4 (ex: STM32F4) pot executa algoritmi de baza cum ar fi urmarirea bazata pe culoare (“color tracking”). Microcontrollerele suporta sisteme de operare minimale, ceea ce inseamna ca algoritmi trebuie dezvoltati si scrisi de la zero de catre programator, cu avantajul de a avea acces direct la functii low-level din firmare, la pini de intrare-iesire (GPIO), timere. Exista si microcontrollere concepute special pentru aplicatii in viziune artificiala (computer vision): CMUcam sau OpenMV .

Sistem de tip SOC (“system on chip”, microcalculatoare) mobile sau tablete sunt de dimensiune un pic mare, au o putere de procesare intre 1-25 GFLOPS si au consumul intre 1 - 6 Watts. Procesoarele mobile ARM pot oferi atat performanta, cat si consum redus. Pentru calcule cu numere complexe (floating point) este recomandata o placa de dezvoltare bazata pe Cortex A9 (ex: ODROID-U3) sau Cortex A15 (ODROID-XU sau Jetson TK1) deoarece hardware-ul specializat pentru operatii cu numere complexe (FPU - floating point unit - coprocesor matematic) este net superior celui din platforme bazate pe ARM Cortex A8. Totusi, solutiile bazate pe Cortex A8 sunt la un pret accesibil (ex: BeagleBone Black) sau ARM11 (ex: Raspberry Pi) si se preteaza bine pentru operatii cu numere intregi. Pentru cea mai buna performanta, sistemul oferit de Nvidia este cea mai buna alegere: aproximativ 192\$ pentru Jetson TK1 care are 192 core-uri grafice bazate pe tehnologia Nvidia CUDA. Daca se doreste o dimensiune fizica redusa, exista Gumstix Overo sau o varianta cu procesor eficient ar fi A20-OLinuXino-LIME2 . Pentru vizualizarea informatiilor si a imaginilor procesate, o alta abordare presupune folosirea telefoanelor mobile sau a tabletelor bazate pe sistemul de operare Android sau Linux. Pentru acces la pini GPIO se poate folosi comunicare prin wireless WiFi sau Bluetooth. Dezvoltarea algoritmilor si a software-ului pentru sisteme bazate pe arhitectura ARM este similara cu cele de pe desktop. Software specializat pentru computer vision este disponibil si pe ARM (ex: biblioteca OpenCV), dar nu este la fel de usor de dezvoltat si ca si sisteme bazate pe x86. Folosind Android sau Linux pentru a executa algoritmi ca si in mod desktop poate insemna ca uneori executia algoritmilor sa fie intrerupti de catre sistemul de operare.

O alta categorie mare este reprezentata de sisteme de tipul SBC (“single board computer), laptop-uri mici sau netbook-uri bazate pe arhitectura x86. Un SBC se poate construi dintr-o placa de baza in format mini ITX (ex: Asus H81i Plus) si un procesor performant (ex: Intel Core i5 sau Intel

Core i7). Avantajul acestor sisteme este dat de puterea mare de procesare (intre 15 - 110 GFLOPS), dar cu dezavantajul de a fi mai mari ca dimensiuni si consum mare de curent (intre 30 - 100 Watt). Pentru vizualizarea rezultatelor un laptop este mai indicat decat un "SBC".

Laptop-uri bazate pe x86 cu placa video dedicata reprezinta o alta categorie mare. Aceste computere ofera o putere foarte mare de procesare (intre 240 - 2200 GFLOPS) dar si un consum de putere pe masura (intre 40 - 110 Watt). Unele notebook-uri ofera placi video dedicate cu suport pentru accelerare GPU CUDA sau OpenCL. Aceste sisteme sunt cel mai usor de programat si de utilizat dar sunt si cele mai costisitoare ca pret de achizitie. Un exemplu ar fi Alienware m14x.

Ultima categorie o reprezinta palcile de tip FPGA (field programmable gate array) care ofera o mare putere de procesare (50 - 1000 GFLOPS) la un consum redus (intre 1 - 3 Watt) si un pret accesibil. Principalul dezavantaj al FPGA este complexitatea sportia pentru design-ul sistemului si al algoritmilor. Acest proces poate dura cate luni sau chiar ani de zile in functie de necesitati si de mana de lucru disponibila. Folosind FPGA, defapt se dezvoltă direct logica hardware-ului de pe placa (chip). Astfel, se dezvoltă un circuit electronic la un nivel foarte jos (low level) care are avantajul de a executa instructiuni constant (este redus riscul de a scade numărul de cadre per secunda). O varianta pentru FPGA ar putea fi Cyclone II Starter Kit împreuna cu un modul de camera video (ex: camera video de 5MP).

In cadrul acestui proiect dimensiunea fizica a sistemului final nu reprezinta o problema. Pentru calcule complexe se vor utiliza sisteme bazate pe x86, dar pentru unele calcule preliminare si pentru controlul mecanismului de tip pan-tilt (care va roti camerele video) este necesara si o platforma de dimensiuni reduse de tip "SOC" (sistem on chip). Pentru acest proiect cea mai buna alegere este utilizarea unei combinatii Arduino/Raspberry Pi 2 pentru interfatare usoara cu motoarele de tip servo sau stepper.

Pe sistemul desktop se vor executa algoritmii de detectie si de urmarire. Folsind accelerare grafica si alte optimizari, un sistem x86 va putea realiza un proces de detectie si urmarire in timp real, dar va putea si trimite comenzi spre sistemul Raspberry Pi prin Ethernet. Rezultatele procesarii imaginilor, mai exact detectia si urmarirea fetei, se vor traduce intr-o serie de comenzi care vor controla mecanismele de tip pan-tilt pe care sunt montate camerele video. Astfel se vor putea urmari persoane care se afla in miscare intr-o incapere sau un spatiu. Daca o persoana iese din cadru, motoarele de tip stepper vor fi actionate astfel incat sa ramana in cadru si sa fie realizata o urmarire continua.

Sistemul de stereoviziune (fig. 2) este format din doua camere video industriale, de tip CCD sau CMOS, monocrome, care vor avea atașata o lentila cu camp vizual mediu sau larg. Acest sistem al camerelor este folosit pentru a extrage informatii despre adancime si urmarirea miscarilor capului sau analiza posturii corpului. Laboratorul de procesare a imaginilor și recunoaștere a formelor al UTCN dispune de mai multe astfel de combinații de camere și lentile, precum și de plăci de achiziție sincronă a imaginilor de la camere, combinația ideală urmând a fi găsită prin experimente.



Fig. 2. Sistemul de camere.

Pe baza informației 3D oferită de sistemul stereo, mecanismul pan-tilt va acționa o a treia cameră. Aceasta camera este color, de mare viteză și va avea un câmp vizual mai îngust pentru a percepe doar zona capului și a estima trasaturile faciale și micro expresiile sau mișcările rapide.

Camera de mare viteză va trebui să ofere un minim de 200 cadre pe secundă (fps = frames per second) și o rezoluție de minim 640x480 pixeli. În general camerele highspeed oferă viteză maximă la o rezoluție mică, pe măsură ce creștem rezoluția scade și numărul de cadre pe secundă care sunt capturate. Această cameră urmează să fie achiziționată în etapa 2.

O variantă posibilă de achiziționat este Optomotive Velociraptor EVO (<http://www.optomotive.com/products/velociraptor-evo>). Aceasta camera este oferită în două variante: cu senzor monocrom și cu senzor color. Ambele folosesc la bază un CMOSIS CMV2000 de dimensiune 2/3", având rezoluția maximă de 2048x1048 pixeli, dimensiunea unui pixel fiind de 5.5 μm, iar imaginile sunt achiziționate la o viteză de 340 cadre/secundă în formatul JPEG. Camera oferă suport standard pentru lentile de tip C mount, iar conectarea la computer se face folosind portul Ethernet.

Camera USB3 Vision, produsă de Ximea (<http://www.ximea.com/en/products/usb3-vision-cameras-xiq-line/mq003mg-cm>) reprezintă o altă opțiune. Camera oferă o viteză de 500 cadre pe secundă la o rezoluție de 640x488 pixeli, iar la o rezoluție de 4 megapixeli suportă o viteză de 90 cadre pe secundă. Există o variantă cu senzor color sau monocrom, iar ambele folosesc la bază CMOSIS CMV300 de dimensiune 1/3". Dimensiunea unui pixel este de 7.4 μm. USB3 Vision suportă formatul de lentile standard de tip C, iar conectarea cu computerul se face folosind o interfață proprietară Ximea sau folosind mufa "opto-isolated" de intrare-iesire.

Megaspeed MS35K, fabricată în Statele Unite ale Americii de către firma Megaspeed reprezintă una dintre cele mai costisitoare camere de mare viteză. Folosește un senzor proprietar produs de Megaspeed de dimensiune 2/3". Rezoluția de achiziție pentru care se obține viteză cea mai mare, 325 cadre/secundă, este de 640x480 pixeli. Rezoluția maximă este de 1280x1020 pixeli pentru care se poate ajunge la maximum de 70 cadre/secundă. Dimensiunea unui pixel este de 7.4 μm, în timp ce camera oferă suport de tip C pentru lentile. Conexiunea cu un computer se face folosind interfața Ethernet.

Tabel 2. Variante posibile de camere de mare viteză, cu pret accesibil proiectului

Model camera	Senzor	Rezoluție optima	Cadre/secunda (la rezoluția optima)	Caracteristici
Optomotive Velociraptor EVO	CMOSIS CMV2000	2048x1048	340	Marime senzor 2/3", pixel = 5.5 μm, conexiune prin Ethernet, pret ~2700€
Ximea USB3 Vision	CMOSIS CMV300	640x488	500	Marime senzor 1/3", pixel = 7.4 μm, pret ~933€
Megaspeed MS35K	Mega Speed B/W CMOS	640x480	325	Senzor 2/3", conexiune Ethernet, pixel = 7 μm, pret ~9300€

Datorită unui raport pret/calitate foarte bun, o alegere potrivită pentru acest proiect o reprezintă camera video oferită de Optomotive. Aceasta camera are un pret avantajos raportat la celelalte soluții, oferă o rată de capturare a imaginilor suficientă pentru acest proiect, dar și o rezoluție mare astfel încât să faciliteze detectia micro-expresiilor sau a gesturilor.

Achiziții efectuate în vederea realizării sistemului multifocal:

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

În vederea realizării sistemului, au fost achiziționate următoarele componente:

- 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.

3. Dezvoltarea metodologiei de calibrare preliminară pentru sistemul senzorial multifocal: Studiul soluțiilor de calibrare existente, elaborarea specificațiilor pentru sistemul de calibrare

Pentru obtinerea informatiei 3D prin procesul de stereoviziune, parametrii implicati in procesul de formare a imaginilor pornind de la punctele lumii reale trebuie cunoscuți. Acești parametri sunt împărțiți în două categorii: parametrii intrinseci, care țin de geometria lentilei și a senzorului de imagine din cameră, și parametri extrinseci, care exprimă legătura dintre camera video și un sistem de coordonate al lumii observate.

Procesul de obținere a parametrilor intrinseci și extrinseci este numit procesul de calibrare a camerelor. Calibrarea parametrilor intrinseci se realizează o singură dată pentru un ansamblu cameră video + lentilă atașată, parametrii obținuți putând fi considerați a fi constanți în timp.

Pentru calibrarea intrinsecă, se pot folosi uneltele de calibrare Caltech (http://www.vision.caltech.edu/bouguetj/calib_doc/). Conform metodologiei, se prezintă camerei un set de șabloane similare cu o tablă de șah (pentru ușurința detectării automate a colțurilor), sub o serie de unghiuri (perspective). O buna practica este de a monta șabloanele pe o suprafață plana și rigidă. Fundalul pe care este listat șablonul, precum și placa pe care este montat trebuie să fie mate și rigide în așa fel încât să nu existe reflexii și imaginea captată să nu prezinte saturații (zone cu luminare la maximul capacității de reprezentare numerică în imagine).

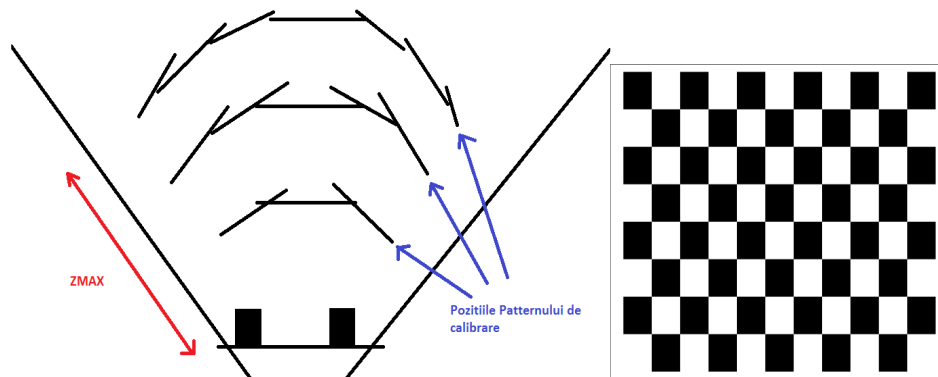


Fig. 3. Model de calibrare, și poziția acestuia față de sistemul stereo.

Se recomandă ca șablonul să fie prezentat camerei pe arcuri de cerc la distanțe variabile astfel încât estimarea distanței focale să fie cât mai precisă. Distanța la care este plasat șablonul trebuie limitată astfel încât această suprafață a modelului în imagine să nu fie mai mică de 1/6 din dimensiunea imaginii. Cu cât distanța la care este plasat șablonul va crește cu atât eroare de măsurare

a punctelor de control din imagine (colțurile șablonului) va crește. Distanța minimă este limitată de distorsiunile introduse de camera și lentile. În figura 3 se observă pozițiile unde trebuie așezat șablonul de calibrare, și un exemplu de model tip tabla de șah.

Pentru calibrarea parametrilor extrinseci se poate folosi calibrarea stereo din pachetul de unelte Caltech, sau se poate folosi o metodă proprie, dezvoltată de grupul de cercetare IPPRG din Universitatea Tehnică, metoda care se bazează pe un număr finit de puncte cu coordonate 3D cunoscute, puncte care se pot identifica ușor în imagine. Pe baza parametrilor intrinseci deja calibrați, și pe baza corespondenței dintre punctele 3D și poziția lor în imagine, se calculează elementele matricei de rotație și a vectorului de rotație folosind metoda de optimizare Gauss Newton.

Pentru calibrarea sistemului multifocal, propunem următoarea strategie:

- Calibrarea individuală a parametrilor intrinseci pentru fiecare cameră video în parte, folosind metoda Caltech.
- Calibrarea parametrilor extrinseci ai sistemului stereo, folosind metoda proprie bazată pe repere 3D. Această calibrare va fi repetată doar dacă există indicii temeinice privind degradarea performanței de reconstrucție, sau dacă se modifică poziția camerelor.
- Calibrarea parametrilor extrinseci ai camerei de viteză mare, pentru mai multe poziții ale sistemului pan-tilt, folosind aceeași metodă ca pentru camerele sistemului stereo.
- Calibrarea automată, în timp real, a camerei mobile, pe baza informației 3D furnizată de sistemul stereo, coroborată cu trăsături invariante extrase din cele trei imagini și puse în corespondență (colțuri, trăsături în mișcare, etc).

4. Evaluarea acurateții algoritmilor de stereoviziune pentru măsurători ale capului și feței

Problema reconstrucției stereo rămâne una dintre cele mai importante subiecte din domeniul viziunii artificiale. Corespondența stereo presupune, așa cum sugerează și numele, găsirea anumitor trasaturi în mai multe imagini care reprezintă aceeași scenă, văzute din perspective diferite. În cazul sistemelor binoculare, după procesul de calibrare și rectificarea imaginilor, trasaturile pot fi cautate pe aceeași linie în ambele imagini. Diferența între pozițiile unde au fost găsite punctele corespondente în cele două imagini se numește disparitate. Valorile disparității pentru toate punctele din imagine constituie harta de disparitate. Odată ce problema corespondenței a fost realizată, distanța până la oricare punct de interes reconstruit din scenă poate fi determinată.

Metodele de găsire a corespondenței stereo (și deci de calculare a disparității) pot fi grupate după rezultatul pe care îl produc. Unele metode produc rezultate dense, încercând să reconstruiască toată scena (dense stereo), altele reconstruiesc scena doar în anumite puncte de interes (sparse stereo).

Metodele sparse caută în ambele imagini anumite trasaturi specifice precum muchii sau colțuri, asadar numărul de puncte reconstruite vor fi mai puține. Acest dezavantaj poate fi contrabalansat de faptul că aceste metode vor fi mult mai rapide, vor consuma mult mai puțină memorie decât metodele dense și viteza de procesare va fi mult mai mare.

Conform studiului prezentat în [1], algoritmii de reconstrucție densă se împart în algoritmi locali și algoritmi globali. Algoritmii locali de reconstrucție sunt mult mai rapizi decât algoritmii globali, dar acuratețea acestor algoritmi este mai scăzută. În algoritmii locali se calculează un scor de potrivire (matching) într-o fereastră, iar calculul disparității se face verificând costul din imaginea de referință cu cel din a doua imagine având în vedere un număr de poziții (disparități) pe care este efectuată căutarea. Disparitatea corectă corespunde costului minim detectat. Metodele globale se bazează pe minimizări de funcții de energie. Aceste metode sunt de regulă foarte costisitoare ca și timp dar produc rezultate foarte precise. Bineînțeles, există și variații la cele două metode menționate

mai sus, din acestea merita sa amintim algoritmi semi-globali de reconstructie stereo care iau putin din cele doua metode cu scopul de a pastra si din acuratetea algoritmilor globali dar si din viteza algoritmilor locali. De asemenea, metodele bazate pe invatare folosind retele neuronale s-au dovedit superioare pe imagini de test din baze de date internationale precum KITTI [2] sau Middlebury [3].

Tabel 3. Metode de calcul a corespondentei stereo

Abordare	Referinte	Descriere succinta
METODE LOCALE		
Block Matching	[4], [5]	Caută un scor de corelare maxim sau o eroare minimă folosind ferestre de matching. Metricile de obtinere de cost presupun fie metrici de rang robuste fie, fie cros-corelare fie compararea regiunilor de intensitate.
Gradient-Based Optimization	[6]	Minimizarea unei functii, spre exemplu sum of squared differences intr-o anumita regiune.
Feature Matching	[7], [8]	Se face matching la trasaturi nu la intensitati. Spre exemplu calcularea unui descriptor census per imagine si efectuarea matchingului asupra descriptorului.
METODE GLOBALE		
Dynamic Programming	[9]	Determină suprafața de disparitate pentru o linie de cautare ca cea mai bună cale între două secvențe de primitive ordonate. Ordinea de regula este data de constrangerea de ordine stereo.
Graph Cuts	[10],[11]	Se determina suprafata disparitatii ca si taietura minima intr-un flux a unui graf.
Nonlinear Diffusion	[12]	Se face o agregare aplicand un process de difuzie.
Belief Propagation	[13]	Se gasesc disparitatile utilizand o transmitere de mesaje intr-o retea

Pasii generali ai unui algoritm de calcul a corespondenței stereo sunt următorii:

1. Calcularea costului de corespondenta intre pixeli din imaginea stanga si dreapta
2. Agregare a spatiului obtinut
3. Calcularea cu sau fara optimizari a disparitatilor
4. Rafinarea hartii de disparitate.

1. Calculul costului de corespondenta

Dat fiind faptul ca in procesul de corespondenta per pixel intre doua imagini se produc erori foarte mari, inaintea inceperii procesului de corelatie de regula se calculeaza un descriptor pe o fereastra in jurul unui pixel, iar comparatia se face folosind rezultatul descriptorului aplicat si o metrica de masurare.

2. Agregare locala a spatiului obtinut

În cazul algoritmilor locali, dat fiind faptul că și corespondența pe baza unui descriptor poate fi eronată, se efectuează o agregare locală unde, în volumul de cost obținut se înlocuiește costul fiecărei disparități cu suma pe o anumită suprafață. Astfel se mediază erorile de matching. În cazul standard se consideră o fereastră patrată de $n \times n$ în jurul unui pixel la o anumită disparitate.

Alte abordări locale efectuează un proces de stereo corespondență pe imaginea normală după care efectuează o fuziune cu harta de disparitate rezultată din corespondența cu imaginea dreaptă după ce, în prealabil a fost efectuată o operație de deformare (warping) a imaginii drepte [14]. Acest warping al imaginii ajută la identificarea suprafețelor care nu sunt fronto-paralele. O soluție care tratează aceeași problemă o găsim prezentată în [15]. Aici autorii rotesc fereastra de matching a descriptorului, adăugând o penalitate la fiecare astfel de warping. Soluția reușește să identifice suprafețele care nu sunt fronto-paralele astfel obținându-se o îmbunătățire de până la 10% pe bazele de date internaționale (kitti) față de metodele locale clasice. Agregarea se efectuează clasic folosind o fereastră de 7×7 .

3. Calcularea Disparităților

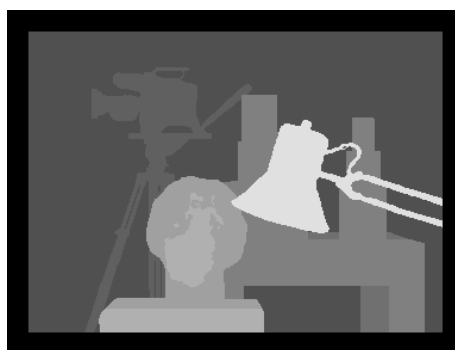
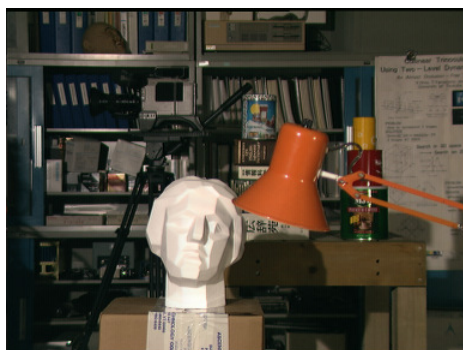
După etapa de agregare se încearcă găsirea disparităților. Acest lucru se realizează printr-un procedeu WTA (winner takes all). Prin acest procedeu se scanează fiecare coloană din volumul de cost obținut și se alege disparitatea cu costul minim. Pentru o acuratețe sporită după găsirea disparității se mai calculează la sub-pixel valoarea disparității. Există multe funcții de calcul a disparității la sub-pixel, printre cele mai bune amintim interpolarea quadratică și echiangularea.

4. Rafinarea

În etapa de rafinare de regulă se realizează un LR consistency check (verificare a consistenței stânga-dreapta), o eliminare a găurilor mici (speckles) din harta de disparitate și o mediere cu 2 filtre mediane (1×9 și 9×1).

Evaluarea algoritmilor de corespondență stereo

Pentru evaluarea algoritmilor stereo se folosesc două benchmark-uri de referință, acceptate de comunitatea internațională [2][3]. Pe aceste benchmark-uri se testează acuratețea algoritmilor de reconstructive. Mai jos am ilustrat două astfel de imagini din baza de date Middlebury. Deși imaginile din baza de date Kitti tratează în general problema reconstrucției stereo pentru scenarii de trafic, se pot folosi aceste imagini pentru testare în cazul în care în imaginile de pe middlebury o anumită problemă nu este așa bine conturată.



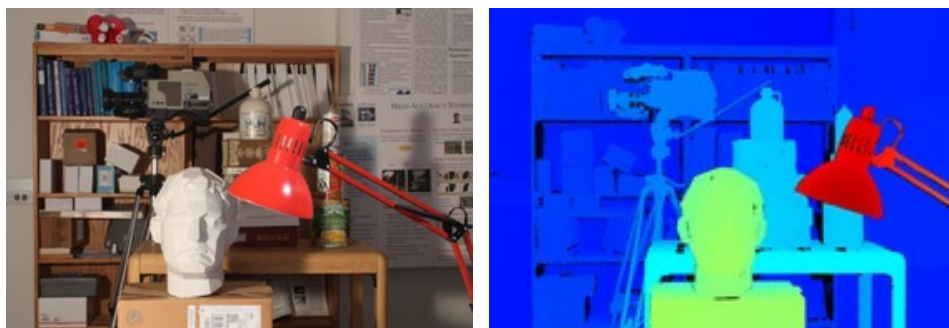


Fig. 4. Imagini de testare ai algoritmilor stereo din baza de date Middlebury.

În continuare, vor fi detaliate câteva din metricile de evaluare care sunt folosite pentru testarea acuratetei algoritmilor stereo:

Out-All – Aceasta eroare măsoară diferența dintre un pixel reconstruit și unul de ground truth. Dacă eroarea este mai mare de un anumit threshold (2 sau 3 pixeli) pixelul reconstruit va fi considerat eronat. Aceasta metodă este folosită cel mai des.

Density – Acest parametru ne informează despre numărul de puncte reconstruite. Adică pentru fiecare punct de ground truth dacă avem un punct reconstruit se consideră că avem densitate pentru respectivul punct. Când se folosește, această metrică, nu ne interesează acuratetea din punctual respective.

Average – Eroarea medie ne spune pentru fiecare pixel reconstruit corect, care este eroarea față de ground truth.

Nonocc – se testează care este eroarea doar în zonele care nu sunt ocluzate, adică zonele care se vad din ambele imagini. Figura de mai jos arată intuitiv zonele la care se referă această metrică pentru primul set de imagini de mai sus. Cu negru este evidențiată zona în care există ocluzii.

Disc – se testează regiunile în apropierea zonelor cu discontinuități.

Toate metricile de mai sus sunt folosite pentru testarea acuratetei procesului de reconstrucție. Un alt parametru necesar evaluării algoritmilor de stereo viziune îl reprezintă timpul necesar reconstrucției. Între timpul de execuție și acuratețe tot timpul există o problemă de pareto-optimal. Acest lucru înseamnă că de cele mai multe ori algoritmi de stereo care se concentrează pe acuratețe au un timp de execuție mai lung. Nu există o metodă de stereo-corespondență care să rezolve ambele probleme (timpul și acuratețea) în procent de 100% (adică un algoritm stereo rapid și de mare acuratețe nu există).

Pentru proiectul nostru va trebui să ne definim și să ne focusăm pe partea care ne interesează cel mai mult.

Mai jos este prezentat mai mult algoritmi de stereo și acuratețea lor corespunzătoare. Unii algoritmi sunt prezentat și pe baza de date de la Kitti unde este evidențiat și timpul lor de execuție.

Algoritmi cei mai rapizi sunt algoritmi locali. Totuși de regulă acești algoritmi prezintă și cea mai proastă acuratețe. În tabelul de mai jos autorii din [15] prezintă comparative diferite metode de corespondență stereo locală. Eroarea prezentată este out all și average.

Tabel 4. Rezultate ale metodelor locale de calcul a corespondenței stereo

Metoda	Rezultate pe diferite imagini				
	Tsukuba	Venus	Teddy	Cones	Average bad pixels
Census	14.0	79.2	21.7	19.4	38.2
CS-CENSUS	25.2	6.59	21.1	19.2	26.1

Metoda	Rezultate pe diferite imagini				
	Tsukuba	Venus	Teddy	Cones	Average bad pixels
MCT	11.2	6.59	21.1	19.2	23.0
MV	12.9	6.53	21.0	19.3	23.3
SWMV	8.47	3.99	14.7	12.1	19.3
SWMCT	11.1	5.38	14.7	12.2	20.2

Ultimele doua solutii ofera rezultate superioare intrucat trateaza problema suprafetelor care nu sunt fronto paralele. De asemenea, ultimele solutii ofera o acuratete de 12% pentru metrica out-all folosind baza de date Kitti.

Un alt algoritm local de mare acuratete si viteza este algoritmul de multi block matching prezentat, evaluat pe baza de date Kitti. Acest algoritm trateaza problema suprafetelor acoperite (occlusions) si efectul de fattening produs in urma agregarii cu ferestre mai mari, pentru a obtine o performanta sporita. Aici avem informatii despre acuratete si performanta algoritmului.

Metoda	Out-Noc	Out-All	Avg-Noc	Avg-All	Density	Timp
MBM	4.35 %	5.43 %	1.0 px	1.1 px	100.00 %	0.2 s

Pe baza de date Middlebury acelasi algoritm va obtine la average bad pixels o valoare de 19.2. De mentionat ca desi acest algoritm are o structura masiv paralela, el nu a fost paralelizat in momentul evaluarii pe Kitti. Daca s-ar folosi orice metoda de paralelizare s-ar obtine rezultate mult mai bune din punctual de vedere al timpului de executie.

Algoritmii locali mai au avantajul ca nu consuma multa memorie si pot fi implementati pe dispozitive embedded. Acesti algoritmi s-ar preta sa fie folositi pentru problema de face and head detection pe dispozitive embedded, intrucat pot reconstrui real time scena cu resurse minime. De asemenea, in momentul alegerii unui algoritm de reconstructive trebuie sa avem in vedere si faptul ca reconstructia reprezinta doar o functie low level, de pre-procesare, in tot pipeline-ul de detectie a fetei.

O clasa de algoritmi imediat urmatoare ca nivel de complexitate este reprezentata de algoritmi semi-globali. Acesti algoritmi rezolva metoda corespondentei folosind programarea dinamica. In functia de energie care trebuie minimizata sunt incluse doua penalitati, una pentru suprafetele inclinate, iar alta pentru ocluzii. Unul din algoritmi de referinta este algoritmul lui Hirschmuler din 2005, Semi-global Matching (SGM).

O implementare de timp real a SGM-ului este rSGM-ul. Aceasta versiune are o acuratete mai scazuta ca si MBM. Timpul de executie este la fel ca si MBM-ul intrucat aceasta versiune de SGM a fost accelerate hardware folosind intrinseci de la intel (SSE).

Method	Out-Noc	Out-All	Avg-Noc	Avg-All	Density	Runtime
rSGM		5.03 %	6.60 %	1.1 px	1.5 px	0.2 s

O alta abordare de SGM care ofera o acuratete putin mai ridicata este prezentata in [18].

Method	Out-Noc	Out-All	Avg-Noc	Avg-All	Density	Runtime
rSGM	4.27 %	5.33 %	1.0 px	1.1 px	100.00 %	5 s

De regula algoritmul SGM ofera o acuratete mai ridicata fata de algoritmi locali, desi in ultima perioada s-au creat algoritmi locali de mai mare acuratete si robuste.

Un dezavantaj major al algoritmilor semi-globali este acela ca consuma mai multa memorie. De asemenea, pentru a putea realiza solutii de timp real trebuie sa folosim metode de accelerare hardware precum GPU sau FPGA.

Metodele globale ofera acuratete mult mai buna decat oricare alte metode, dar avand in vedere ce spune principiul pareto, timpul de executie este mult mai ridicat. Aceste metode nu pot fi folosite la solutii de timp real si de regula consumul de memorie este foarte ridicat.

Un astfel de algoritm este „adaptive belief propagation” [16]. Acest algoritm de gasire a corespondentei foloseste segmentarea bazata pe culoare in imaginea referinta si un scor care se auto adapteaza si care are scopul de a maximiza numarul de corespondente corecte. Structura scenei este modelata de un set de suprafete planare care sunt estimate folosind o tehnica robusta. In loc sa-i fie asignata o valoare de disparitate fiecarui pixel, un plan de disparitate este asignat fiecarui segment. Planul optim al disparitatii e aproximat folosind belief propagation. Rezultatele experimentale pe baza de date Middlebury evidentiaza acuratetea metodei, dar timpul de executie al metodei este de 25 de secunde pe un calculator cu un procesor de 2.21 Ghz, fara paralelizare.

Din rezultatul analizei facute deducem ca nu exista o metoda “cea mai buna” care sa rezolve toate problemele. Dupa realizarea sistemului senzorial multifocal, se vor experimenta mai multi algoritmi si se va alege cel care ofera cel mai bun raport dintre acuratete si timp de executie.

Bibliografie

- [1] D. Scharstein, R. Szelinski “A taxonomy and evaluation of dense two frame stereo correspondence algorithms”, IJCV 7, 1/3, 2002.
- [2] <http://www.cvlibs.net/datasets/kitti/>
- [3] <http://vision.middlebury.edu/stereo/eval3/>
- [4] O. Faugeras, B. Hotz, H. Matthieu, T. Vieville, Z. Zhang, P. Fua, E. Theron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy, "Real Time Correlation-Based Stereo: Algorithm, Implementations and Applications," INRIA Technical Report 2013, 1993.
- [5] D. N. Bhat and S. K. Nayar, "Ordinal Measures for Image Correspondence," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, pp. 415-423, 1998.
- [6] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Proc. Int'l Joint Conf. A. I., pp. 674-679, 1981.
- [7] S. Birchfield and C. Tomasi, "Multiway Cut for Stereo and Motion with Slanted Surfaces," Proc. Int'l Conf. Computer Vision, vol. 1, pp. 489-495, 1999.
- [8] F. Bigone, O. Henricsson, P. Fua and M. Stricker, “Automatic Extraction of Generic HouseRoofs from High Resolution Aerial Imagery,” Proc. European Conf. Computer Vision, pp. 85-96, 1996.
- [9] P. N. Belhumeur, "A Bayesian Approach to Binocular Stereopsis," Int'l J. Computer Vision, vol. 19, no. 3, pp. 237-260, 1996
- [10] Y. Boykov, O. Veksler and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222-1239, 2001.
- [11] Y. Boykov and V. Komolgorov, “An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision,” Proc. Third Int'l Workshop on Energy
- [12] D. Scharstein and R. Szeliski, "Stereo Matching with Non-Linear Diffusion," Int'l J. Computer Vision, vol. 28, no. 2, pp. 155-174, 1998.
- [13] J. Sun, H. -Y. Shum and N. -N. Zheng, “Stereo Matching Using Belief Propagation,” Proc. European Conf. Computer Vision, pp. 510-524, 2002.
- [14] N. Einecke and J. Eggert: Stereo Image Warping for Improved Depth Estimation of Road Surfaces. IV 2013.
- [15] M. Muresan, S. Nedeveschi, M. Negru Improving local stereo algorithms using binary shifted windows, fusion and smoothness constraint, ICCP 2015
- [16] Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information Heiko Hirschmuller, CVPR, 2005
- [17] A. Klaus, M. Sormann and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. ICPR 2006.

[18] Z. Zhong: Efficient Learning based Semi-Global Stereo Matching. 2015 submitted.

5. Proiectarea modelelor pentru geometria și poziția capului, și pentru trăsăturile faciale: Studiul modelelor și a tehnicilor de urmărire existente pentru cap și trăsături faciale

Studiul expresiilor faciale a fost o preocupare constantă a oamenilor încă din perioada Antichității (sec. V î.Hr.) când s-a dezvoltat fiziognomonica, disciplină care încearcă să determine caracterul unui om pe baza caracteristicilor sale faciale. În secolul XIX s-au publicat două studii importante care au o legătură directă cu metodele actuale de analiză a expresiilor faciale. Neurologul francez Guillaume Duchenne de Boulogne a publicat un studiu [1] în care a analizat modul în care mușchii faciali generează expresii, prin utilizarea unor stimuli electrici pentru a genera contracții musculare faciale și capturarea unor imagini cu expresiile generate. Ulterior, Charles Darwin a scris un tratat [2] în care a analizat modul general în care apar expresiile, atât la oameni cât și la animale, și le-a clasificat în mai multe categorii. Un punct de cotitură în studiul expresiilor faciale a fost la începutul anilor 1970, când Paul Eckman și colegii săi au început studiul expresiilor faciale pentru a determina emoțiile ascunse ale oamenilor; acestea au un impact major în dezvoltarea sistemelor actuale de recunoaștere automată a expresiilor faciale.

Un subiect controversat în psihologie este teza universalității expresiilor faciale. Teza a fost inițiată de Darwin, care a susținut că expresiile faciale implicate în exprimarea emoțiilor de bază sunt universale [2]. Această teorie a fost intens dezbătută. În prezent s-a ajuns la un consens, în care se afirmă că interpretarea expresiilor faciale pentru expresiile de bază (fericire, tristețe, mânie, surprindere, dezgust și frică) este constantă în diferitele culturi, dar apariția acestor depinde de contextul social, dat fiind faptul că diferitele culturi au impus anumite reguli în ceea ce privește exprimarea expresiilor (display rules).

Fața umană poate exprima mii de expresii faciale diferite, dar nu toate acestea sunt corelate cu emoțiile resimțite. Expresiile faciale se pot clasifica în:

- semnale conversaționale (conversational signs), utilizate pentru a accentua limbajul vorbit;
- embleme (facial emblems), cum ar fi ridicarea unei sprâncene în semn de scepticism sau „trasul cu ochiul”;
- manipulatori faciali (facial manipulators), de exemplu mișcarea buzelor;
- expresiile emoționale, care pot fi sincere și ne-sincere [3].

Paul Eckman a identificat următoarele indicii la nivel facial pentru a depista înșelăciunea:

- microexpresiile;
- expresiile înăbușite (squelched);
- mușchii faciali siguri (reliable muscles);
- asimetriile expresiilor;
- anumiți parametri legați de dinamica expresiilor (durata expresiei, coordonarea cu discursul vorbit și cu mișcările corporale).

Microexpresiile sunt considerate unele dintre cele mai sigure surse de detecție a înșelăciunii. Acestea sunt definite ca expresii faciale scurte, cu o durată cuprinsă între 1/12 și 1/25 secunde, care apar atunci când oamenii încearcă să-și ascundă sentimentele, fie ca o formă de represie (ascundere intenționată), fie ca o formă de refulare (ascundere inconștientă). Expresiile înăbușite sunt expresii întrerupte care durează mai mult decât o micro-expresie: odată cu apariția expresiei, persoana în cauză devine conștientă de începutul acesteia și o suprimă, de regulă prin altă expresie facială, cum ar fi un zâmbet.

Eckman a descoperit de asemenea că nu toți mușchii faciali pot fi la fel de ușor controlați: de exemplu, foarte puține persoane sunt capabile să-și îndrepte colțurile gurii în jos fără a-și mișca vârful bărbiei, dar totuși această mișcare apare în expresiile de tristețe. Astfel, prin cunoașterea și

urmărirea acestor mușchi faciali siguri (reliable) se pot determina emoțiile. De asemenea, s-a descoperit că expresiile veritabile, sincere sunt simetrice, în timp ce expresiile mimate, false prezintă anumite asimetrii.

Sistemele de recunoaștere automată a expresiilor și a micro-expresiilor își găsesc aplicații interesante într-o varietate de domenii: psihologie, înțelegerea proceselor sociale și a unor tulburări psihice, animații grafice, interacțiunea om-calculator, detecția oboselii șoferilor, marketing, estimarea durerii etc.

Metodologii pentru măsurarea expresiilor faciale

Electromiografia (electroneuromiografia, EMG) este o metodă care permite măsurarea activității musculare prin plasarea unor electrozi de suprafață sau a unor ace-electrod pe piele și înregistrarea descărcării electrice a țesutului muscular. Această metodă permite determinarea imediată a contracțiilor musculare și poate înregistra mișcări de mică amplitudine care nu sunt vizibile cu ochiul liber. EMG este o procedură intruzivă, care, deși este tolerată bine de majoritatea pacienților, produce un anumit disconfort. În plus, subiecții sunt conștienți de măsurarea reacțiilor lor, și astfel rezultatele testelor pot să fie neconcludente. Un alt dezavantaj este faptul că măsurătoarea se efectuează asupra unei regiuni și nu asupra mușchilor și nu s-a determinat dacă această metodă permite diferențierea clară între diferitele expresii [4].

Începând cu anii 1960, studiul expresiilor faciale ca modalitate de exprimare a emoțiilor a devenit o direcție de interes în domeniul psihologiei. Deși acest domeniu era studiat cu mult timp înainte, încă nu exista un sistem parametrizat care să permită analiza obiectivă a expresiilor faciale. Prin urmare, accentul s-a pus mai ales pe dezvoltarea unor metodologii care să le permită cercetătorilor să clasifice orice mișcare facială și astfel să decidă asupra stării emoționale a subiectului observat ([5], [6]).

Cel mai utilizat sistem pentru măsurarea expresiilor faciale pe baza mișcărilor musculare este sistemul FACS (Facial Action Coding System) dezvoltat de Paul Eckman pe baza unui studiu al psihologului Hjortsjo [7]. Metodologia FACS permite măsurarea oricărei expresii faciale care poate fi exprimată de oameni, prin descompunerea acesteia la nivelul acțiunilor fundamentale ale unor mușchi sau grupe de mușchi (Action Unit, AU), și la intervalul temporal în care a apărut expresia. Intensitatea unui Action Unit (gradul de activare al mușchilui) poate fi indicată prin calificativele: A (Trace - urmă), B (Slight - superficial), C (Pronounced - pronunțat), D (Extreme - extrem) și E (Maximum – maxim).

Pornind de la metodologia FACS s-au dezvoltat și alte sisteme, cum ar fi EMFACS (Emotional Facial Action Coding System) și FACSAID (Facial Action Coding System Affect Interpretation Dictionary), dar acestea consideră doar mișcărilor faciale care sunt implicate în exprimarea emoțiilor.

Dezvoltarea unui sistem automat capabil să recunoască și să clasifice mișcărilor feței pe baza metodologiei FACS ar avea un impact semnificativ într-o mulțime de domenii: de la înțelegerea comportamentului uman la viziunea artificială în care se urmărește recunoașterea expresiilor faciale, de la animații grafice la psihiatrie și medicină.































Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Fig. 5. Exemple de Action Units definite în metodologia FACS.

La începutul anilor 1990, în domeniul graficii pe calculator a apărut aceeași problemă ca și în comunitatea psihologică înainte de introducerea metodologiei FACS: deși animația fețelor era un domeniu de mare interes, nu exista nici un standard unificator și fiecare sistem folosea parametri diferiți pentru animarea fețelor. Standardul internațional Face and Body Animation (FBA) (ISO/IEC 14496-1 & -2) ([8], [9]) a fost dezvoltat de Motion Experts Group pentru a facilita dezvoltarea unei metodologii standard de exprimare a mișcărilor corpului și ale feței. FBA definește un model facial într-o stare neutră pe care se definesc 84 de puncte de interes (feature points, FP) (Figura 6), a căror poziție poate fi utilizată pentru a recunoaște anumite expresii faciale sau pentru a anima fețele. De asemenea, standardul definește 68 de parametri de animație ai feței (Facial Animation Parameters, FAP) care efectuează diferite deformări ale feței din starea neutră. Valoarea unui FAP indică magnitudinea deformării aplicate modelului din starea neutră. Pentru ca interpretarea FAP-urilor să se realizeze într-un mod consistent pe orice model al feței, acestea sunt exprimate prin Face Animation Parameter Units (FAPU) – valori definite de distanța dintre diferite trăsături faciale și exprimate de distanța dintre diferite FP. Exemple de FAPU sunt distanța dintre ochi ES sau dimensiunea gurii. Fiecare FAP este definit de FP-urile asupra cărora acționează, de direcția de mișcare și de FAPU care este folosit ca unitate a mișcării.

FAP-urile pot fi grupate în 66 de trăsături de nivel jos și 2 trăsături de nivel înalt: expresii și viseme (imagini faciale specifice utilizate pentru a descrie un anumit sunet). Sunt definite cele 6 expresii primare: bucurie, mânie, tristețe, frică, deznădejde și surprindere.

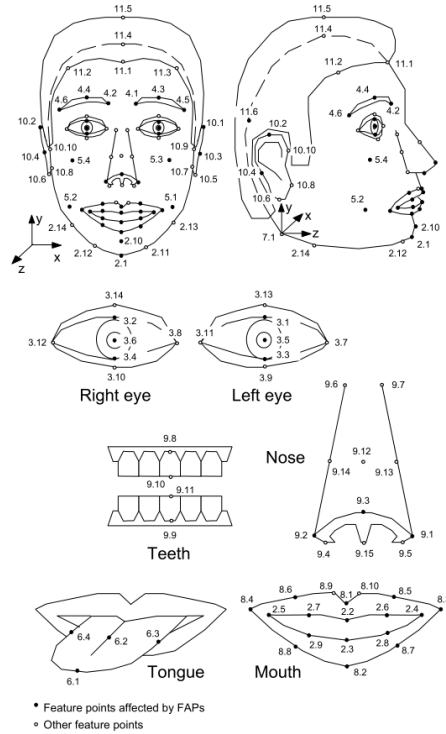


Fig. 6. Feature points definite în standardul MPEG4 (din [8])

Există mai multe studii care au realizat corespondența între AU din metodologia FACS și FAP din standardul MPEG4 [10].

Modelul Candide 3 [11] este un model dezvoltat la începutul anilor 90 pe baza metodologiei FACS (7d), fiind compatibil și cu standardul FBA dezvoltat de MPEG-4.

Candide 3 (Figura 7) este compus dintr-un număr redus de vârfuri (112), ceea ce permite generarea și reconstrucția rapidă a modelului cu resurse computaționale reduse. Fiecare vârf are o corespondență directă cu un FP definit de standardul FBA. Nu toate FP-urile sunt reprezentate în modelul Candide: de exemplu, nu sunt definite părul, urechile, limba și dinții. Modelul Candide 3 definește 12 parametri prin care se poate deforma modelul neutru al feței către o față specifică, numiți Shape Units (SU). Astfel, modelul generat poate reprezenta cel puțin fețele cu o formă comună. Exemplu de SU sunt: înălțimea capului, poziția sprâncenelor, poziția verticală a nasului etc. Modelul include mai multe Animation Units. Un Animation Unit este definit prin implementarea corespunzătoare a unuia sau mai multor AU în modelul Candide. Cu alte cuvinte, Shape Units descriu forma particulară a feței subiectului, iar Animation Units sunt deformări ale formei neutre.

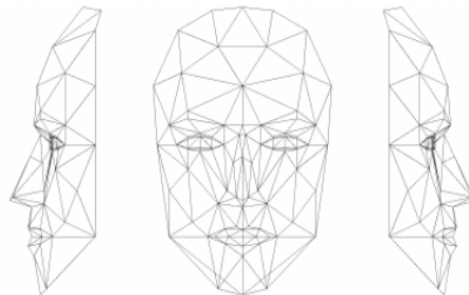


Fig. 7. Modelul Candide 3 (din [11])

Deși în literatura de specialitate s-au propus modele faciale mult mai complexe [12], [13]), care conțin mii de poligoane pentru a descrie forma feței și care sunt bazate pe modele anatomice complexe, modelul Candide este încă des folosit în domeniul viziunii artificiale, datorită simplității sale, care permite procesarea și recunoașterea expresiilor faciale cu o complexitate redusă. De exemplu, Face Tracking SDK al senzorului Microsoft Kinect utilizează modelul Candide 3 pentru a urmări fața oamenilor în timp real.

Stadiul actual în domeniu și contribuții originale posibile

1. Urmărirea, clasificarea și analiza mișcărilor oculare

Urmărirea poziției ochilor este un domeniu intens studiat, iar o acoperire exhaustivă a lucrărilor publicate în această direcție este aproape imposibilă. Pentru un studiu amănunțit al sistemelor de eye tracking și ale aplicațiilor acestora se pot consulta lucrările [14] și [15].

Pe baza sursei de lumină utilizată, sistemele de urmărire a ochilor pot fi clasificate în metode active (care utilizează LED-uri IR) și metode pasive (care se bazează pe lumina naturală). Metodele active se bazează pe o proprietate fizică a pupilei care își modifică aspectul în imaginea capturată în funcție de poziția sursei de lumină IR și facilitează astfel procesul de detecție.

Din punctul de vedere al echipamentului hardware utilizat, sistemele de urmărire a ochilor pot fi montate pe cap sau amplasate pe o pereche de ochelari ușori sau la distanță.

În funcție de abordarea conceptuală utilizată în procesul de detecție, sistemele de urmărire ale ochilor se pot grupa în următoarele clase: metode bazate pe formă, metode bazate pe aspect și metode hibride (care combină primele două clase pentru a valorifica beneficiile acestora). Metodele bazate pe formă se bazează pe un model predefinit al formei ochilor și utilizează o măsură de similaritate pentru a detecta ochii în imaginile de test, în timp ce metodele bazate pe aspect [16] detectează ochii în funcție de distribuția de culoare sau de răspunsul unor filtre într-o zonă din jurul ochilor.

În pofida numărului mare de metode propuse pentru urmărirea ochilor, încă se mai pot aduce îmbunătățiri în acest domeniu.

Principalele contribuții originale care preconizăm că vor fi aduse în dezvoltarea acestui modul sunt:

- Sistemul dezvoltat nu va fi amplasat pe capul subiectului (nu va fi head-mounted), fiind un sistem non-invaziv. Astfel, subiectul nu va fi conștient de faptul că este analizat și nici nu va resimți vreun inconvenient în acest proces. Sistemele head-mounted sunt dificil de utilizat mai ales în cazul copiilor, care au tendința de a le îndepărta.

- Nu se vor utiliza surse de lumină în infraroșu (IR). Majoritatea sistemelor de urmărire ale ochilor se bazează pe surse de lumină IR, deoarece acestea nu sunt vizibile ochiului uman și facilitează procesul de detecție. Principalul dezavantaj al acestor metode este faptul că nu pot fi utilizate în mediul exterior, datorită interferențelor legate de iluminare.

- Rezultatele acestui modul vor fi corelate cu alte informații legate de orientarea capului, prezența unor ocluziuni, poziția corpului, expresii faciale etc. Astfel, sistemul dezvoltat va fi mult mai tolerant față de ocluziuni, cum ar fi ochelarii sau condițiile de iluminare. În plus, prin combinarea acestor informații se va obține o mai bună înțelegere a proceselor emoționale și cognitive ale oamenilor.

- Rezultatele obținute vor fi analizate pentru a determina diferiți parametri: dimensiunea pupilei în diferite momente de timp, direcția de privire, frecvența clipirilor, distanța inter-pupilară etc. Acești parametri au o importanță deosebită pentru o mulțime de aplicații cum ar fi: optometria, determinarea gradului de dificultate cognitivă al unei sarcini (cognitive workload), detecția oboselii etc.

- Folosind o cameră high-speed se vor detecta și se vor determina parametri sacadelor: amplitudinea, viteza maximă (peak velocity), durata. Analiza acestor mișcări are un impact important atât în medicină, cât și în înțelegerea anumitor procese psihologice. Mai multe studii medicale au arătat că anumite diferențe în mișcările ochilor pot indica prezența anumitor tulburări de neuro-dezvoltare, cum ar fi: deficitul de atenție, autismul, sindromul alcoolismului fetal sau al unor boli neurologice.

2. Estimarea poziției capului

În ultimele decenii, estimarea poziției capului a fost intens studiată ([17], [18], [19]), datorită numeroaselor aplicații posibile pe care le are într-o varietate de domenii. În [17] se propune o taxonomie bazată pe metodele conceptuale utilizate în estimarea poziției capului. Metodele bazate pe aspect (active appearance methods) utilizează o bază de date cu imagini de test (adnotate cu o orientare discretă a poziției capului) și compară imaginile de test cu această bază de date pentru a găsi orientarea cea mai apropiată. Metodele bazate pe tablouri de detecție (detector arrays methods) utilizează mai mulți clasificatori, fiecare dintre aceștia fiind specializat pe o anumită orientare a capului, și poziția capului se determină pe baza detectorului cu cel mai mare scor. Alte metode utilizează regresia neliniară (nonlinear regression methods) pentru a găsi o potrivire funcțională între imagini și orientarea capului. Metodele flexibile (flexible methods), cum ar fi AAM (Active Appearance Methods) sau ASM (Active Shape Methods), potrivesc un model deformabil pe o nouă imagine, iar orientarea capului este estimată pe bază comparațiilor la nivel de trăsături cu cele stocate într-o bază de date. Metodele geometrice (geometric methods) se bazează pe locația diferitelor elemente faciale (ochi, vârful nasului, gura etc.) pentru a estima poziția capului pe baza poziției lor relative. Sistemele bazate pe urmărire (tracking) determină modificările globale ale poziției capului pe baza mișcării observate (optical flow) între mai multe cadre succesive. În sfârșit, metodele hibride combină una sau mai multe din procedeele descrise mai sus pentru a compensa dezavantajele lor. În ultima perioadă, senzorii de adâncime au devenit accesibili din punct de vedere al prețului (cum ar fi senzorul Microsoft Kinect sau ASUS Xtion) și s-au dezvoltat mai multe metode care combină informațiile de culoare cu cele de adâncime pentru a elimina multe din problemele specifice imaginilor [19].

Deși s-au efectuat o mulțime de studii în ceea ce privește estimarea poziției capului, încă mai există numeroase provocări care trebuie să fie depășite și nici unul din sistemele existente nu a reușit să le depășească pe toate. Sistemul ideal ar trebui să fie complet autonom (să nu fie necesară o fază de inițializare manuală sau să presupună o anumită orientare a capului la începutul estimării). În al doilea rând, algoritmi de urmărire a capului ar trebui să fie capabili să estimeze un interval continuu de orientări în timp real, și în plus ar trebui să funcționeze atât pentru imagini de apropiere cât și pentru imagini de depărtare, indiferent de rezoluția acestora. De asemenea, poziția capului ar trebui să poată fi estimată corect și precis în orice condiții, chiar și când fața nu este orientată către cameră.

Principalele contribuții originale care vor fi aduse de acest modul sunt:

- Dezvoltarea unui metode precise de estimare a poziției capului bazată pe stereo-viziune și pe un model simplu 3d al capului (modelul Candide [11]).
- Pentru urmărirea robustă a capului se vor folosi algoritmi de estimare probabilistică (filtru de particule), ceea ce va permite urmărirea în timp real și revenirea rapidă din situațiile în care capul nu mai este vizibil în cadrele video. De asemenea, prin utilizarea acestui cadru de estimare probabilistică nu va fi nevoie de o fază de inițializare a procesului de tracking.
- Sistemul senzorial original utilizat nu va impune nici o constrângere de mișcare asupra subiectului urmărit.

Rezultatele modulului de urmărire a poziției capului vor fi corelate cu cele ale poziției ochilor pentru a determina direcția de privire. Aceste informații au un impact semnificativ în crearea unor

interfețe utilizator de ultima generație: deja există anumite aplicații care le permit utilizatorilor să controleze mouse-ul doar cu mișcările capului. Sistemele de urmărire a capului de ultimă generație vor influența de asemenea și industria automobilelor: din poziția capului și estimarea direcției de privire se vor putea determina automat situațiile în care șoferul nu este atent la drum. De exemplu, s-a determinat că utilizarea telefoanelor mobile în timpul condusului reprezintă cauza pentru 18% dintre accidentele mortale din Statele Unite ale Americii.

3. Analiza și recunoașterea expresiilor faciale

Analiza expresiilor faciale și a dinamicii are un impact major în educație, medicină, securitate, mai ales în zilele noastre în care culturile eterogene se întrepătrund, iar comunicarea non-verbală poate fi o sursă de înțelegere reciprocă. În plus, analiza expresiilor faciale și a amplitudinii acestora poate fi utilizată ca o metodă de evaluare sau diagnosticare non-intruzivă a unor maladii care implică anumite modificări la nivelul feței: depresie, boala Parkinson, sindromul Moebius.

Recunoașterea automată a expresiilor faciale se poate diviza în trei module importante: detecția/urmărirea fețelor, extragerea trăsăturilor și clasificarea expresiilor. Primul pas în recunoașterea expresiilor este detecția feței; la ora actuală, cel mai utilizat algoritm în această fază este metoda Viola Jones [16] pentru orientările frontale ale fețelor. După detecția feței, pentru partea de tracking se poate urmări o mulțime simplă de trăsături faciale, cum ar fi ochii, sau trăsături mai detaliate, pe baza unor modele cum ar fi Active Appearance Models, metoda Lukas-Kanade sau Eigentracking [20]. După ce fața a fost detectată și diferitele puncte de pe față au fost urmărite, urmează un proces de normalizare pentru a compensa orientarea capului și pentru a alinia geometric toate trăsăturile feței. Din imaginile normalizate se extrag trăsăturile pe baza cărora se va face recunoașterea expresiilor; aceste trăsături pot fi bazate fie pe proprietăți geometrice (informații legate de forma și poziția elementelor faciale: ochi, buze, vârful nasului), fie pe aspectul feței, folosind informații legate de textura feței. În final, se utilizează un algoritm de inteligență artificială pentru a clasifica fiecare expresie facială pe baza trăsăturilor extrase.

Încă există multe aspecte legate de recunoașterea automată a expresiilor faciale care pot să fie îmbunătățite: aceste sisteme trebuie să fie complet automate, să funcționeze chiar și în cazurile în care nu este vizibilă întreaga față și să fie invariante la schimbările de iluminare și la machiaj, ochelari, păr facial etc. [21]

Principalele contribuții care vor fi aduse de acest model și potențialele aplicații ale lor sunt:

- Investigarea posibilității de a utiliza mișcările și expresiile faciale ca o metodă de identificare biometrică;
- Segmentarea precisă a feței și localizarea trăsăturilor faciale. Rezultatele obținute vor fi integrate într-o aplicație de recomandări pentru alegerea coafurii sau ochelarilor în funcție de forma feței. În anul 1991, Claude Juillard a redefinit termenul de visagism, adaptându-l pentru și opticienii, prin utilizarea relației dintre forma feței și mimica persoanei.

4. Detecția, analiza și clasificarea micro-expresiilor

Recunoașterea automată a expresiilor faciale este un domeniu care a fost intens studiat în ultimii 30 de ani în domeniul viziunii artificiale. Deși micro-expresiile sunt din ce în ce mai studiate pentru a înțelege 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 [22], CASME [23] și CASME II [24] conțin expresii spontane. Baza de date SMIC conține secvențe video capturate cu o cameră de mare viteză de 100

FPS cu 164 de micro-expresii de la 16 participanți; micro-expresiile sunt clasificate în două categorii (pozitive și negative). 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 controlate cu o cameră 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 precizi de urmărire și detecție a mișcării.

Recunoașterea automată a micro-expresiilor are trei faze distincte: segmentarea regiunilor faciale care vor fi analizate, extragerea unor trăsături pe baza cărora se vor recunoaște micro-expresiile și recunoașterea propriu-zisă a acestora folosind algoritmi din domeniul inteligenței artificiale.

Prima etapă a procesului de recunoaștere a microexpresiilor presupune stabilirea zonei faciale care va fi analizată; pentru această etapă fața fie este împărțită în mai multe segmente dreptunghiulare în jurul trăsăturilor faciale mai importante (ochi, buze, frunte) [25], [26], [27], fie se utilizează un model deformabil pentru a partiționa fața în mai multe regiuni de interes [28], [29]. Pentru fiecare regiune selectată a feței se definesc Action Unit-urile care acționează în acea zonă.

În a doua etapă a procesului de recunoaștere se extrag anumiți descriptori spațio-temporali care vor fi utilizați în descrierea transformărilor faciale apărute de-a lungul timpului. Acești descriptori se bazează pe extragerea fluxului optic (optical flow) dens în regiunile selectate [28], [27], pe descriptori de textură [29], sau pe histograma 3D a gradientilor [25], [26].

Ultimul pas în recunoașterea automată a microexpresiilor presupune utilizarea unor algoritmi de învățare automată (supervizată [29] sau ne-supervizată [25], [25]) pentru a recunoaște fiecare microexpresie.

Recent s-au dezvoltat mai multe sisteme de recunoaștere automată a micro-expresiilor, dar în acest domeniu există încă o mulțime de aspecte care trebuie să fie îmbunătățite. Sistemele viitoare care se vor dezvolta ar trebui să fie mai tolerante cu privire la mișcarea capului: toate sistemele dezvoltate până la ora actuală presupun că poziția capului nu se modifică. În ceea ce privește rezultatele modulelor de clasificare: acestea fie raportează microexpresiile apărute ca fiind pozitive sau negative, fie le clasifică în una din cele 6 categorii de emoții de bază. Din punct de vedere aplicativ ar fi mult mai util ca aceste sisteme să raporteze ce modificări au apărut la nivelul Action Unit-urilor și analiza efectivă să fie realizată de psihologi în funcție de aceste modificări. De asemenea, pentru detecția înșelătoriei și a emoțiilor ascunse ar fi util să se determine și diferiți parametri ai micro-expresiilor: onset (momentul în care expresia începe să apară), apex (momentul de maximă amplitudine), offset (momentul în care expresia dispare), simetria și amplitudinea acestora.

Contribuțiile originale care vor fi aduse de acest modul sunt:

- Investigarea posibilității de a utiliza un sistem hardware special format din două camere normale declanșate la momente de timp apropiate pentru a obține o frecvență a cadrelor mai mare (TODO)

- Dezvoltarea unui sistem în care sunt identificate mișcărilor faciale (Action Unit-urile) pe baza metodologiei FACS și nu doar clasificarea micro-expresiilor (pozitive /negative sau în una din cele 6 emoții considerate universale). Aceasta ar avea impact într-o mulțime de domenii: de exemplu, în identificarea depresiei și în studiul efectelor injectării cu botox. Metodele actuale de diagnosticare a depresiei se bazează în principal pe un interviu clinic a pacientului și a familiei acestora. Analiza expresiilor faciale ar putea reprezenta un viitor mijloc eficient și sistematic de diagnoză al acestei

maladii prin utilizarea informațiilor comportamentale, care sunt indicatori clari ai unor probleme psihologice.

- Analiza parametrilor micro-expresiilor: onset (momentul în care expresia începe să apară), apex (momentul de maximă amplitudine), offset (momentul în care expresia dispare) și a amplitudinii acestora.

- Dezvoltarea unui sistem mai tolerant fata de mișcările capului și capabil să detecteze micro-expresii și în cazuri în care sunt prezente și alte expresii faciale.

Micro-expresiile au o importanță crucială în securitate (incluzând lupta împotriva terorismului internațional), înțelegerea comportamentului uman și detectarea emoțiilor ascunse (Eckman a descoperit micro-expresii de anxietate și teamă prin urmărirea unei înregistrări a unei paciente care a încercat să se sinucidă în care aceasta zâmbea mult și părea foarte fericită). La ora actuală în Statele Unite ale Americii în unele aeroportului este implementat programul SPOT (Screening Passangers by Observation Techniques), în care angajați ai aeroportului sunt antrenați în a observa pasagerii cu comportament suspect prin analiza micro-expresiilor și prin conversație. Totuși, unele studii arată că oamenii, chiar și dacă au fost antrenați în a detecta micro-expresiile obțin o rată de detecție redusă (puțin peste 50%). Astfel, putem concluziona că dezvoltarea unui sistem automat de detecție a micro-expresiilor este foarte necesară.

Bibliografie

- [1] Guillaume Duchenne, *Mécanisme de la Physionomie Humaine*. Paris: Jules Renouard, 1862.
- [2] Charles Darwin, *The Expression of the Emotions in Man and Animals*. United Kingdom: John Murray, 1872.
- [3] Paul Eckman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton & Company, 2009.
- [4] Paul Eckman, "Facial expressions of emotion: An old controversy and new findings.," *Philosophical Transactions: Biological Sciences*, vol. 335, no. 1273, 1992.
- [5] Paul Eckman and William Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.
- [6] Carrol Izard, *The Maximally Discriminative Facial Movement Coding System (MAX)*. Newark: University of Delaware Instructional Resource Center, 1979.
- [7] Carl Hjortsjo, *Man's Face and Mimic Language.:* Studen litteratur, 1969.
- [8] Visage Tehnologies. (2015, Nov.) <http://www.visagetehnologies.com/>. [Online]. <http://www.visagetehnologies.com/uploads/2012/08/MPEG-4FBAOverview.pdf>
- [9] Igor Pandzi and Robert Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation And Applications.:* John Wiley & Sons, Inc., 2003.
- [10] A Raouzaiou, N Tsapatsoulis, K Karpouzis, and S Kollias, "Parameterized facial expression synthesis based on MPEG-4," *EURASIP Journal on Applied Signal Processing*, vol. 2002, 2002.
- [11] Jörgen Ahlberg, "CANDIDE-3 - An Updated Parameterised Face," 2001.
- [12] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition.," in *IEEE 6th International Conference on Advanced Video and Signal Based Surveillance*, 2009.
- [13] Volker Blanz and Thomas Vetter, "A Morphable Model for the synthesis of 3d faces," in *26th annual conference on Computer graphics and interactive techniques*, New York, 1999.
- [14] Dan Witzner, Ji, Qiang Hansen, "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, 2010.
- [15] A. T. Duchowski, "A Breadth-First Survey of Eye Tracking Applications," *Behavior research methods, instruments, computers*, vol. 34, no. 4, 2002.

- [16] Paul Viola and Michael Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vision*, vol. 57, no. 2, 2004.
- [17] Erik Murphy-Chutorian and Mohan Manubhai Trivedi, "Head pose estimation in Computer Vision – a survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, 2009.
- [18] Błażej Czapryński and Adam Strupczewski, "High Accuracy Head Pose Tracking Survey," *Active Media Technology*, vol. 8610, 2014.
- [19] G Fanelli, J. Gall, and L. van Gool, "Real Time 3D Head Pose Estimation: Recent Achievements and Future Challenges," in *5th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2012.
- [20] F. De la Torre and J. F. Cohn, "Facial Expression Analysis," in *Guide to Visual Analysis of Humans: Looking at People.*: Springer, 2011.
- [21] Vinay Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," *CoRR*, 2012.
- [22] Xiaobai Li, T. Pfister, Xiaohua Huang, Guoying Zhao, and M Pietikainen, "A Spontaneous Micro-expression Database: Inducement, collection and baseline," in *Automatic Face and Gesture Recognition (FG)*, 2013.
- [23] Wen-Jing and Wu, Qi and Liu, Yong-Jin and Wang, Su-Jing and Fu, Xiaolan Yan, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Automatic Face and Gesture Recognition (FG)*, 2013.
- [24] Wen-Jing Yan et al., "CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation," *PLOS One*, 2014.
- [25] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *3rd International Conference of Crime Detection and Prevention*, 2009.
- [26] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta, "Facial Micro-Expression Detection in Hi-Speed Video Based on Facial Action Coding System (FACS)," *IEICE TRANSACTIONS on Information and Systems*, vol. E96, no. 1, 2013.
- [27] S Godavarthy, D Goldgof, S Sarkar M Shreve, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011.
- [28] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, and Guoying Zhao, "A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition," *IEEE Transactions On Affective Computing*, vol. 99, 2015.
- [29] Thomas Pfister, Li Xiaobai, Zhao Guoying, and Michael Pietikainen, "Recognising spontaneous facial micro-expressions," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

Cluj Napoca, 25.11.2015

Conf. Dr. Ing. Radu Danescu