DETECTAREA AUTOMATĂ A FEȚELOR UMANE. METODA VIOLA-JONES

Mihnea Horia Vrejoiu mihnea@dossv1.ici.ro

Anca Mihaela Hotăran ahotaran@ici.ro

Institutul Național de Cercetare-Dezvoltare în Informatică - ICI București

Rezumat: Departe de a mai reprezenta o tehnologie exotică, desprinsă din povestiri sau filme ştiinţifico-fantastice, detecția automată a fețelor în imaginile digitale a devenit deja parte a vieții noastre cotidiene. Astăzi, practic orice utilizator al unei camere foto obişnuite poate remarca apariția automată a unor dreptunghiuri colorate care încadrează figurile subiecților umani asupra cărora este îndreptat obiectivul respectivei camere. Puțini sunt totuşi cei care s-au întrebat cum este posibil aceasta şi mai puțini probabil, sunt cei care cunosc mai mult sau mai puțin răspunsul la această întrebare. Articolul de față îşi propune să prezinte problematica detecției automate a fețelor, cu o succintă trecere în revistă a dificultăților specifice, precum şi a principalelor tipuri de abordări în soluționarea acesteia, cu detalierea - la un nivel care se doreşte cât mai accesibil posibil - a uneia dintre cele mai cunoscute şi utilizate metode, şi anume Viola-Jones. Articolul nu se doreşte a fi un studiu exhaustiv sau o monografie a domeniului, ci numai o introducere în acesta, cât mai pe înțelesul celor mai mulți, încercând totodată să reliefeze frumusețea şi eleganța inovației, eficiența şi performanța aduse de metoda Viola-Jones.

Cuvinte cheie: detecția fețelor, vedere artificială, analiză de imagini, învățare automată, caracteristici de tip Haar, cascadă atențională, AdaBoost, clasificator, imagine integrală.

Abstract: Far from representing an exotic techology, drawn from sci-fi stories or movies, the automatic face detection imperceptibly became already part of our everyday life. In nowadays, almost any user of an average photo camera could see the automatic drawing of a colored rectangle which is framing the faces of human subjects towards which the respective camera objective is pointed to. Few of them asked themselves however how is this possible, and probably fewer know more or less the answer to this question. The present paper aims to present the problematics of automatic face detection, with a brief overview of the specific difficulties, and of the main types of approaches in solving this problem, while detailing - at a level which aims to be as accessible as possible - of one of the most well known and used method, the Viola-Jones one. This paper don't intend to be an exhaustive survey or a monography on the domain, but only an introduction to this one, easily understandable by most people, while also trying to highlight the beauty and elegance of the innovation, the efficiency and performances brought by the Viola-Jones method.

Keywords: Face Detection, Computer Vision, Image Analysis, Machine Learning, Haar-like features, Attentional cascade, AdaBoost, classifier, integral image.

1. Introducere

Departe de a mai reprezenta o tehnologie exotică, desprinsă din povestiri sau filme ştiinţifico-fantastice, detecția automată a fețelor în imaginile digitale a devenit deja parte a vieții noastre cotidiene.

Astfel, în zilele noastre, practic oricine a avut ocazia să utilizeze o cameră foto digitală obişnuită (sau chiar camera foto a unui telefon mobil ieftin) şi a putut observa pe ecranul acesteia cum, la încadrarea subiecților umani, apare (câte) un dreptunghi / pătrat colorat care încadrează figurile respective. Această funcție de localizare automată a fețelor permite o mai bună încadrare, precum şi stabilirea automată a parametrilor de focalizare şi expunere corectă a fețelor respective, prin comparație cu restul scenei. Probabil însă că puțini sunt aceia care s-au întrebat cum anume reuşeşte camera foto să distingă faptul că într-o anumită regiune a imaginii există o figură umană iar în altele nu, respectiv că o anumită configurație de pixeli reprezintă o față, în timp ce alte configurații nu. Şi asta practic în timp real! Cum camerele foto sau telefoanele mobile, cel puțin cele de clasă medie, ieftine, nu sunt dotate cu cipuri cu putere foarte mare de procesare, înseamnă că algoritmul respectiv şi implementarea sa trebuie să fie extrem de eficiente. În plus, software-ul de detecție a fețelor integrat în camerele foto obişnuite a evoluat, ajungând astăzi nu numai să localizeze figurile umane, dar chiar să şi identifice apariția zâmbetului pe acestea şi să comande declanşarea automată în momentul respectiv.

De asemenea, detecția automată a fețelor reprezintă astăzi o componentă fundamentală ca prim pas în aplicațiile din ce în ce mai variate şi răspândite de recunoaştere a fețelor, precum şi în cele vizând noi modalități evoluate de interacțiune om-calculator. Ea joacă totodată un rol extrem de

important şi în ceea ce înseamnă etichetarea automată a volumului uriaş de date de tip imagine disponibile pe Internet, într-o nouă paradigmă a acestuia, prin asocierea unor metadate descriptive utile pentru clasificarea lor, stabilirea de legături categoriale şi/sau semantice între ele şi alte date disponibile şi, în cele din urmă, pentru căutarea şi regăsirea acestora grupate pe criterii variate.

Totuşi, trebuie precizat faptul că problema detectării figurilor umane în imagini nu este tocmai una trivială, datorită varietăţii uriaşe în care acestea pot apărea în percepţia senzorului 2-D, atât din cauza trăsăturilor şi particularităţilor fizionomice, culorii, dimensiunilor, poziţiei, acoperirii parţiale, fundalului complex, dar mai cu seamă din cauza zonelor de lumină şi umbră determinate de poziţia şi distribuţia sursei / surselor de lumină.

O definiţie pentru problema detecţiei feţelor ar fi următoarea [1]: fiind dată o imagine (digitală) arbitrară, să se detecteze toate figurile (dacă există) şi să se stabilească localizarea (poziţia şi dimensiunea) lor exactă.

2. Abordări şi soluţii în domeniul detecţiei automate a feţelor

De-a lungul timpului au existat numeroase tipuri de abordări, ca metode, tehnici şi algoritmi, pentru găsirea unor soluţii optime, fiabile, performante şi eficiente, de tratare a problemei detectării figurilor umane în imagini (digitale). Primul sistem de detectare a feţelor a fost dezvoltat în anii '70 [3]. La începutul anilor '90, dezvoltarea tehnicilor de recunoaştere a feţelor a făcut necesară şi dezvoltarea unor algoritmi mai performanţi pentru detectarea acestora, ca prim pas în recunoaşterea automată. Detectarea feţelor poate fi privită ca un caz particular al detectării claselor de obiecte, presupunând localizarea acestora în imagini (digitale) indiferent de orientare / poziţionare, condiţii de iluminare, scalare, particularităţi şi expresie facială. Domeniului i s-a acordat o atenţie deosebită în ultimii 15 ani, în principal datorită creşterii numărului de aplicaţii comerciale şi din domeniul legal care necesită autentificarea personală (de exemplu, controlul accesului – care presupune şi recunoaşterea feţei, supravegherea spaţiilor publice, interacţiunea om-calculator etc.) pe de o parte şi, pe de altă parte, dezvoltarea unui număr mare de dispozitive de captură de imagini ieftine.

Deşi pentru vederea umană pare o sarcină banală, detectarea automată a feţelor este o problemă dificilă, practic imposibil de definit / precizat şi tratat complet şi riguros, varietatea modurilor posibile în care pot apărea figurile umane în imagini fiind practic infinită. Şi asta nu atât din cauza trăsăturilor şi particularităţilor fizionomice şi a dimensiunilor variate ale acestora în diferite imagini, cât mai ales din cauza varietăţii în percepţia lor în 2-D, datorate poziţiei (orientării / înclinării, scalării) efective în imagine şi în special efectelor jocului luminii şi al umbrelor asupra reliefului acestora, precum şi din cauza contextului înconjurător variat şi imprevizibil.

Abordările şi soluţiile încercate de-a lungul timpului au trebuit să-şi propună să facă faţă tuturor situaţiilor care pot fi întâlnite, indiferent de particularităţile subiectului şi în orice context de poziţionare, dimensiune, scalare, încadrare, fundal, sau iluminare s-ar afla acesta. O altă necesitate pentru majoritatea, datorită specificului aplicaţiilor, a fost să producă rezultate în timp real. Unele au reuşit mai bine, altele mai puţin bine. Performanţa sistemelor de detecţie a feţelor este dată de două componente: rata / procentul fals-pozitivelor (confuziilor - raportări false ca feţe ale unor regiuni din imagini care nu conţin în realitate o faţă), precum şi rata / procentul fals-negativelor (rateurilor - regiuni ale imaginilor care în realitate conţineau o faţă, dar care nu au fost raportate ca atare). Pentru un sistem cât mai general, fiabil şi robust, acestea trebuie să fie ambele cât mai mici, ideal tinzând spre zero fiecare.

O observaţie care trebuie făcută, este aceea că detecţia feţelor, indiferent de metoda utilizată, nu este şi nu poate fi o ştiinţă exactă. Aşa cum chiar oamenii pot fi păcăliţi de imagini 2-D care par a conţine o figură umană deşi în realitate nu există niciuna acolo, la fel şi algoritmii de detecţie a feţelor pot fi induşi în eroare în anumite situaţii. Acest fenomen poartă denumirea de „*pareidolia*” în limba engleză [2] şi se referă la recunoaşterea aparentă a ceva semnificativ (de obicei o figură sau siluetă umană) în ceva care nu-l conţine în mod natural. Există numeroase exemple în acest sens, poate cel mai spectaculos fiind „figura” de pe Marte – o fotografie luată în regiunea Cydonia de pe Marte ce pare a conţine o figură umană sculptată în stâncile de pe solul marţian respectiv, sau imaginea Sfintei Fecioare „descoperită” de o americancă într-un sandwich cu brânză prăjit, sau

chiar nenumăratele aparente figuri formate de configurațiile noroase sau ale frunzelor copacilor în jocul de lumină şi umbră. Aşa cum am mai spus, software-ul de detecție a fețelor poate fi de asemenea păcălit de astfel de situații, deci se poate vorbi despre o rată de fals-pozitive (detecții de fețe unde nu sunt) şi una de fals-negative (nedetectarea unei fețe acolo unde de fapt exista una) ale acestora.

Putem enumera următorii factori principali care pot induce probleme în detectarea automată a fețelor (în imagini bidimensionale):

- poziția şi orientarea acestora în imagine (frontal, profil, sub un unghi etc.) - anumite caracteristici faciale (ochi, nas) putând fi parțial sau total ascunse;
- prezența / absența unor componente structurale - unele caracteristici faciale precum barbă, mustață, ochelari putând fi, sau nu, prezente şi existând o mare variabilitate a acestora din punct de vedere al formei, culorii sau dimensiunilor;
- expresia facială - geometria feței fiind afectată de aceasta;
- obturarea - fețele putând fi parțial mascate (acoperite) de alte obiecte (inclusiv alte fețe);
- condițiile în care a fost realizată fotografia - iluminarea (spectrul, poziția şi/sau distribuția sursei / surselor de lumină, intensitatea) şi caracteristicile aparatului foto (lentilele, senzorul) afectând foarte puternic felul în care o figură apare în imagine.

În general, metodele utilizate, indiferent de tipul acestora, sunt laborioase, presupun numeroase iterații, analize pe diferite criterii, scalări, filtrări, comparații, şi evident au solicitat eforturi şi timp pentru a fi puse la punct. Unele se bazează pe antrenări anterioare ale unor clasificatoare (în general cu două clase: față şi non-față) de diferite tipuri, utilizând rețele neuronale, *support vector machines* (SVM), modele ascunse (*hidden*) Markov etc., în timp ce altele încearcă să se bazeze pe cunoştințe şi observații *a priori* codate programatic şi utilizate astfel la detecție, iar altele încearcă o mixare între aceste două tipuri de abordări în diferite etape de analiză. În general, toate metodele recurg la multiple operații de scalare şi/sau normalizare fie a imaginii, fie a unei subferestre de analiză, prin glisare peste aceasta, ori a modelelor / şabloanelor sau caracteristicilor a căror prezență este verificată în imaginea țintă.

Algoritmii de detecție a fețelor pot fi împărțiți în patru mari categorii, unele metode fiind localizabile la granițele dintre acestea [1][3]:

- ***Metoda bazată pe cunoştințe***

Aceasta se bazează pe cunoştințele despre geometria tipică a feței umane şi dispunerea spațială a caracteristicilor faciale. Metodele bazate pe cunoştințe folosesc reguli pentru a descrie forma, textura şi/sau alte caracteristici ale elementelor faciale (cum ar fi ochi, nas, bărbie, sprâncene, etc.). Ele utilizează în general o abordare ierarhică, prin care se examinează imaginea la diverse rezoluții. La nivelul cel mai înalt sunt găsiți posibili candidați utilizând o descriere grosieră a geometriei feței. La nivelurile inferioare se extrag caracteristicile faciale şi regiunile din imagine sunt identificate ca făcând parte dintr-o față sau nu, pe baza unor reguli prestabilite implicând caracteristicile faciale şi poziționarea lor (relativă). Principala problemă a acestei tehnici este aceea de a găsi o modalitate eficientă de transpunere a cunoştințelor umane despre geometria feței în reguli bine definite. Dacă regulile sunt prea detaliate (stricte) sistemul nu va permite detectarea fețelor care nu trec toate regulile. Dacă regulile sunt prea generale vor produce foarte multe fals-pozitive. O altă problemă este legată de faptul că metoda nu funcționează corespunzător în condiții de iluminare variabilă şi pentru diferite orientări ale capului, deoarece este imposibil să se prevadă toate cazurile posibile.

- ***Abordarea bazată pe caracteristici invariante***

Această metodă încearcă să găsească unele caracteristici structurale comune indiferent de condițiile de iluminare sau de unghiul din care a fost capturată imaginea. Au fost utilizate diverse caracteristici structurale: caracteristici faciale locale, textura, forma şi culoarea pielii. Caracteristicile faciale locale, cum ar fi ochii, sprâncenele, nasul, gura etc., sunt extrase utilizând

filtre multi-rezoluție sau de tip derivativ, detecția contururilor, operații morfologice, sau folosind anumite praguri. Pe baza acestora se construiesc modele statistice care descriu relații între caracteristici şi se verifică existența feței. De asemenea, ca instrumente pentru verificarea candidaților au fost utilizate rețelele neuronale, potrivirea / corelarea de grafuri (*matching*), sau arbori de decizie. În cazul imaginilor color, este foarte utilă folosirea culorii pielii ca element de detecție inițială a candidaților posibili, segmentarea imaginilor după schema de culoare fiind rapidă computațional şi robustă la schimbarea unghiului de captură a imaginii, la scalare, umbrire, sau în prezența unui fundal complex. Alte tehnici folosesc o combinație de caracteristici pentru a creşte acuratețea detecției, cum ar fi de exemplu utilizarea texturii, formei şi culorii pielii pentru a găsi candidați şi apoi a caracteristicilor faciale locale (ochi, nas, gură) pentru a-i verifica / valida. Această metodă este însă ineficientă dacă imaginea este coruptă sau (foarte) deformată.

- ***Metoda bazată pe modele / şabloane (templates)***

Metoda detectează întâi capul, care este aproximativ eliptic, folosind filtre, detectoare de contur sau siluete. Sunt apoi extrase contururile caracteristicilor faciale utilizând cunoştințe despre geometria feței (modele / şabloane predefinite manual sau parametrizate de o funcție). Se calculează corelația dintre caracteristicile extrase din imagine şi modele / şabloane predefinite ale caracteristicilor faciale. Metodele bazate pe modele / şabloane predefinite sunt sensibile la scalare şi variația formei şi poziției. Pentru a rezolva această problemă au fost propuse modele / şabloane deformabile, care modelează geometria feței cu modele elastice care permit translatarea, scalarea şi rotația. Modelul utilizează şi parametri corespunzători informației de intensitate.

- ***Metoda bazată pe apariție / înfățişare (appearence)***

Această metodă utilizează un număr mare de exemple corespunzătoare diferitelor tipuri de variații. Detectarea feței este privită ca o problemă de recunoaştere de forme cu două clase: „față" şi „non-față". Sunt utilizate tehnici de analiză statistică şi învățare automată pentru a descoperi proprietățile statistice sau funcția de distribuție a probabilității *pattern*-urilor intensității pixelilor din imaginile corespunzătoare celor două clase. Cele mai utilizate metode din această categorie pentru detectarea fețelor sunt: *eigenfaces*, LDA, rețele neuronale, *support vector machines* (SVM) şi modele ascunse (*hidden*) Markov. Aceste metode pot fi înțelese într-un context probabilistic. O imagine sau un vector de caracteristici rezultat dintr-o imagine este o variabilă aleatoare x, caracterizată de funcțiile de densitate p(x|față) şi p(x|non-față). Pentru a clasifica subimaginile, se foloseşte un clasificator Bayesian sau de asemănare maximă (*maximum likelihood*). Din păcate, o implementare directă a unui clasificator Bayesian nu este posibilă datorită dimensionalității lui x. O altă abordare este descoperirea unei funcții de discriminare (de exemplu: suprafețe de decizie, hiper-planuri de separare, etc.) între clase. În mod obişnuit, se reduce dimensionalitatea imaginii şi apoi se utilizează o astfel de funcție de discriminare (bazată de exemplu pe distanță) sau o suprafață neliniară, calculată cu rețele neuronale multistrat.

3. Metoda Viola-Jones

În anul 2001, Paul Viola şi Michael Jones [4] au prezentat un cadru nou, inovativ, pentru detectarea obiectelor arbitrare în imagini, pe care l-au rafinat pentru detecția fețelor. Algoritmul este cunoscut drept metoda Viola-Jones şi este unul dintre cei mai robuşti, performanți, eficienți şi utilizați, reprezentând practic un punct de cotitură în dezvoltarea aplicațiilor practice de timp real utilizând detecția fețelor, cum este cazul celor incluse în camerele foto digitale de astăzi.

Metoda Viola-Jones oferă o viteză remarcabilă şi o rată foarte înaltă de acurateţe – cercetătorii raportând o rată de fals-negative (nedetecție) de sub 1% şi una de fals-pozitive de sub 40%, chiar când sunt utilizate doar cele mai simple filtre. Metoda completă utilizează după cum vom arăta mai jos, până la 38 de filtre sau clasificatoare.

Principala inovație adusă de Viola şi Jones a fost aceea de a nu încerca să analizeze direct imaginea în sine, ci anumite „caracteristici" (*features*) dreptunghiulare în aceasta. Aceste caracteristici, inspirate, printr-o analogie cu analiza formelor de undă complexe din sistemul orto-normal de funcții Haar de bază, sunt cunoscute sub denumirea de „caracteristici de tip Haar"

(*Haar-like features*), după matematicianul ungur Alfred Haar care a trăit şi creat la începutul secolului 20. Trebuie făcută totuşi – pentru evitarea posibilelor confuzii – precizarea că nu este vorba aici de *wavelet*-uri Haar.

În primul rând, dacă imaginea de analizat este una color, aceasta este transformată într-una în nivele de gri, în care apar doar nivelele de strălucire / luminanţă, informaţia de culoare fiind neglijată. Este de subliniat aici faptul că independenţa de culoare conferă metodei o mare generalitate. În practică se poate utiliza pentru fiecare pixel (x, y) al imaginii o relaţie de forma:

$$i(x, y) = 0{,}299\ R(x, y) + 0{,}587\ G(x, y) + 0{,}114\ B(x, y),$$

unde valorile R, G, B reprezintă respectiv componentele de roşu, verde şi albastru ale valorii respectivului pixel (x, y) în spaţiul RGB utilizat frecvent în reprezentarea digitală a imaginilor color, pe 3 sau 4 octeţi (24 sau 32 de biţi). Valorile sunt cuprinse între 0 şi 255 (cât poate fi reprezentat pe un octet).

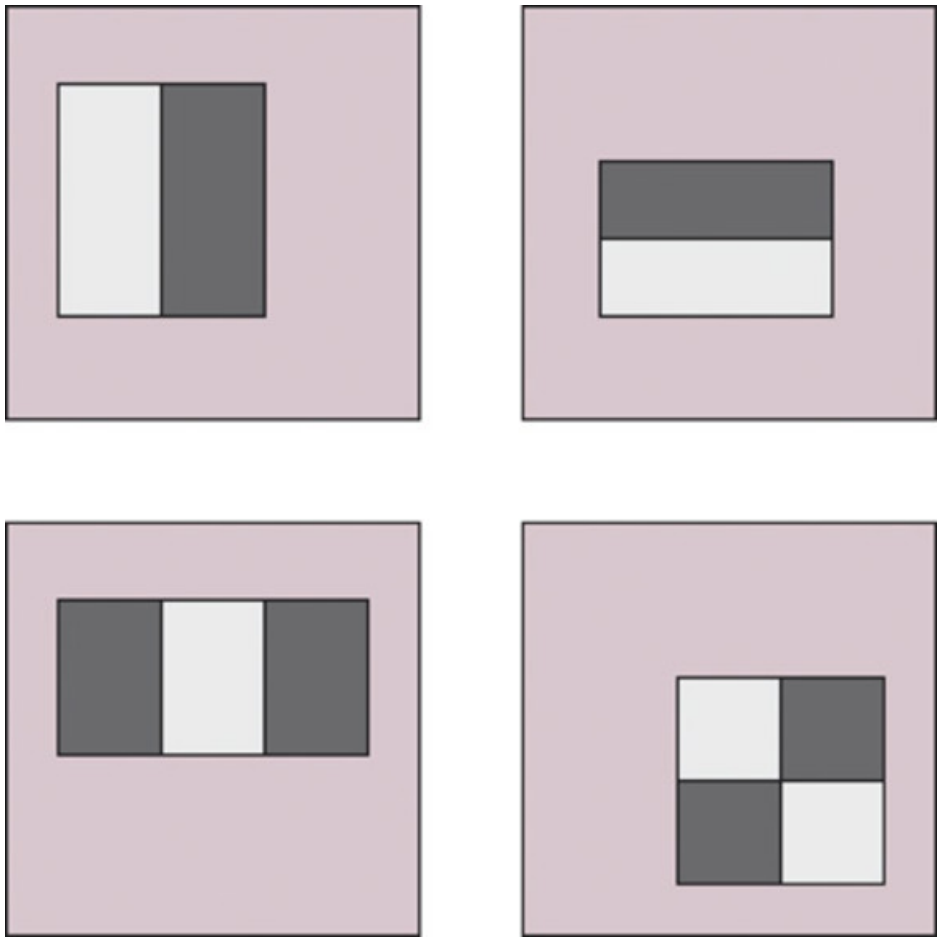

Figura 1. Câteva tipuri de caracteristici definite de Viola şi Jones

Pe imaginea în 256 de nivele de gri astfel obţinută pot fi analizate anumite caracteristici dreptunghiulare prin sumarea valorilor intensităţilor pixelilor în diferite blocuri dreptunghiulare. Astfel, pot fi detectate în imagine caracteristici formate din blocuri mai întunecate, adiacente unor blocuri mai luminoase, suma pixelilor din primele fiind mai mică decât cea a pixelilor din cele din urmă. Trebuie precizat faptul că aceste caracteristici nu reprezintă totuşi în niciun fel anumite caracteristici / trăsături faciale specifice.

Viola şi Jones au definit mai multe tipuri de astfel de caracteristici, reprezentate prin câte două, trei, sau patru blocuri dreptunghiulare adiacente, cu tentă întunecată şi respectiv deschisă. În Figura 1 sunt reprezentate câteva dintre acestea.

Trebuie precizat faptul că într-o anumită caracteristică, toate blocurile dreptunghiulare componente au fiecare aceeaşi formă şi aceleaşi dimensiuni. Aceste caracteristici pot fi evaluate la orice scală sau poziție în imagine, cele reprezentate în Figura 1 având dimensiuni arbitrare, doar cu titlu de exemplu. Fiecărei astfel de caracteristici îi este asociată o valoare care este calculată ca diferența între suma pixelilor din regiunea de imagine delimitată de dreptunghiurile deschise, minus suma pixelilor din dreptunghiurile închise care compun caracteristica respectivă. Valoarea caracteristicii este apoi utilizată de un filtru pentru a determina dacă acea caracteristică este prezentă sau nu în imaginea originală.

Pare că mecanismul descris mai sus ar fi destul de costisitor din punct de vedere computațional, adică destul de lent. O a doua inovație spectaculoasă şi importantă adusă de metoda Viola-Jones constă într-o optimizare computațională ingenioasă, pentru îmbunătățirea vitezei de sumare a intensităților pixelilor din blocurile dreptunghiulare componente ale caracteristicilor definite. Astfel, pentru fiecare imagine originală de analizat se generează inițial o aşa numită „imagine integrală", mapată 1 la 1 peste imaginea originală, în care fiecare punct capătă valoarea sumei tuturor pixelilor din imaginea originală situați la stânga şi deasupra coordonatelor punctului respectiv, inclusiv.

Valorile tuturor punctelor ii(x, y) din imaginea integrală pot fi calculate cu o singură trecere prin imaginea inițială, pornind din colțul din stânga-sus (x = 0 şi y = 0), pixel cu pixel, linie după linie, de sus în jos.

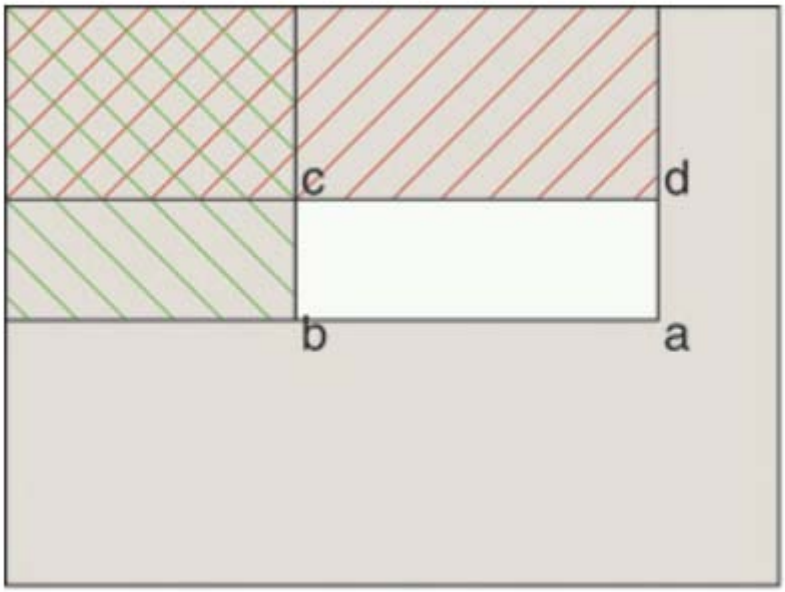

Figura 2. Calculul sumei intensităților pixelilor dintr-o regiune dreptunghiulară a imaginii originale utilizându-se valorile punctelor din imaginea integrală

Fie, pentru oricare linie y din imaginea originală:

$$s(x, y) = s(x-1, y) + i(x, y), \quad \text{cu convenția: } s(-1, y) = 0,$$

suma cumulativă a valorilor tuturor pixelilor i(x, y) din linia y până în poziția x inclusiv.

Valoarea oricărui punct (x, y) din imaginea integrală poate fi calculată ca:

$$ii(x, y) = ii(x, y-1) + s(x, y), \quad \text{cu convenția: } ii(x, -1) = 0.$$

Termenul „integrală" are aceiaşi semnificație ca în definirea matematică a unei integrale, respectiv aria de sub o curbă, obținută prin sumarea unor arii dreptunghiulare elementare.

Odată calculată imaginea integrală pentru fiecare punct corespunzător din imaginea originală, suma intensităților pixelilor din oricare dreptunghi arbitrar din aceasta din urmă poate fi calculată cu uşurință. Astfel, pentru calcularea sumei tuturor pixelilor din dreptunghiul abcd din imaginea

originală (Figura 2), pot fi utilizate numai valorile punctelor a, b, c şi d corespondente din imaginea integrală, efectuându-se numai trei operaţii simple (două scăderi şi o adunare) între acestea, astfel:

$S_{abcd} = ii_a - ii_b - ii_d + ii_c$.

În acest mod pot fi calculate extrem de eficient din punct de vedere computaţional valorile asociate caracteristicilor definite, la orice scală şi în orice poziţie în imaginea originală. Dar ce şi cum se poate obţine pe baza acestora? Prin comparaţie cu analiza directă a intensităţii pixelilor, caracteristicile oferă o vedere mai grosieră, în rezoluţie scăzută, a imaginii, fiind potrivite pentru caracterizarea unor particularităţi locale în imagine prin detectarea limitelor între regiuni luminoase şi întunecoase, dungi / bare şi alte structuri simple.

Ce au realizat mai departe Viola şi Jones a fost să implementeze un sistem de antrenare / învăţare. Au utilizat ca *input* pentru rutina de detecţie a feţelor un set de imagini de 24 x 24 de pixeli conţinând feţe şi un alt set de astfel de imagini de 24 x 24 pixeli care nu conţineau feţe şi au antrenat rutina să recunoască feţele şi să elimine non-feţele. Au fost utilizate aproape 5.000 (circa 4.900) de astfel de exemple conţinând feţe şi 10.000 de exemple de non-feţe, colectate arbitrar de pe Internet.

În ce a constat însă de fapt învăţarea? Utilizând imagini 24 x 24, există circa 45.000 de moduri diferite de a plasa una dintre cele patru tipuri de caracteristici prezentate în Figura 1 pe o astfel de imagine (sau peste 160.000 în cazul tuturor tipurilor de caracteristici definite de Viola şi Jones). De exemplu, pentru primul tip de caracteristică, pot fi considerate dreptunghiuri de 1 x 2 pixeli, până la 1 x 24, apoi 2 x 2 până la 2 x 24 şi aşa mai departe. Aceste caracteristici de diverse dimensiuni pot fi plasate în poziţii diferite pe imagine astfel încât să fie testate toate caracteristicile posibile, de toate dimensiunile posibile, în fiecare poziţie posibilă.

Se poate observa imediat că numărul caracteristicilor posibile, de circa 45.000 (sau 160.000), este de departe mai mare ca numărul de pixeli dintr-o imagine 24 x 24, respectiv 576, deci este evident că trebuie redus cumva numărul celor care sunt utilizate. Să ne amintim că pentru fiecare caracteristică se calculează diferenţa între sumele pixelilor din regiunile deschise şi respectiv închise ale acesteia. Se poate stabili un prag pentru aceste diferenţe (care poate fi ajustat în timpul antrenării) pe baza căruia o caracteristică să fie considerată ca detectată sau nu. Utilizându-se acesta, se aplică fiecare dintre cele 45.000 (respectiv 160.000) de caracteristici posibile la setul de învăţare.

S-a dovedit însă, că anumite caracteristici nu sunt utile în determinarea faptului că o imagine reprezintă o figură sau nu, respectiv că nu există nicio corelaţie în modul în care o caracteristică identifică o faţă şi respectiv nu o identifică şi reciproc. La aceste caracteristici s-a renunţat. Pe de altă parte, alte caracteristici s-au dovedit a avea o rată de succes mare în eliminarea subferestrelor de tip non-faţă şi aici intervine practic învăţarea.

Viola şi Jones au făcut o serie de experimente cu caracteristicile rămase pentru a determina cea mai bună metodă de utilizare a acestora pentru clasificarea unei imagini ca faţă sau non-faţă. În cele din urmă au decis să utilizeze o variantă a unui sistem de învăţare automată (*machine learning*) denumit AdaBoost (de la Adaptive Boosting), pentru a construi un clasificator.

AdaBoost este o tehnică de Inteligenţă Artificială (I.A.) similară reţelelor neuronale, un meta-algoritm adaptiv formulat iniţial de Zoav Freund şi Robert Schapire [9], dezvoltat pentru a combina caracteristici slabe într-un clasificator mai puternic. Fiecărei caracteristici dintr-un clasificator îi este ataşată o pondere (ajustată în timpul învăţării) care defineşte precizia clasificatorului. Ponderi mici înseamnă caracteristici slabe, în timp ce ponderi mari sunt asociate cu caracteristici puternice. Dacă suma ponderilor caracteristicilor care au răspuns pozitiv pe o anumită imagine depăşeşte un anumit prag (ajustabil de asemenea în timpul învăţării), se decide că imaginea respectivă este o faţă.

Un alt element important care trebuie subliniat este acela că Viola şi Jones au utilizat AdaBoost nu numai pentru antrenarea clasificatorului, dar şi pentru selectarea unui număr considerabil mai mic dintre toate caracteristicile iniţial definite, care să fie utilizate în cele din urmă.

Viola şi Jones au constatat în cursul învăţărilor că există două caracteristici care, combinate şi

ajustate corespunzător de AdaBoost într-un singur clasificator, pot recunoaşte 100% feţele, cu o rată a fals-pozitivelor de 40% (respectiv, 60% dintre non-feţe sunt rejectate de acest clasificator).

În Figura 3 este prezentat sugestiv acest tip simplu de clasificator în acţiune. El utilizează două caracteristici pentru a testa imaginea: o caracteristică orizontală, care măsoară diferenţa între regiunea mai întunecată a ochilor şi cea mai luminată a pometilor obrajilor şi caracteristica cu trei dreptunghiuri verticale, care testează regiunile mai întunecate ale ochilor cu regiunea mai luminoasă a şeii nazale.

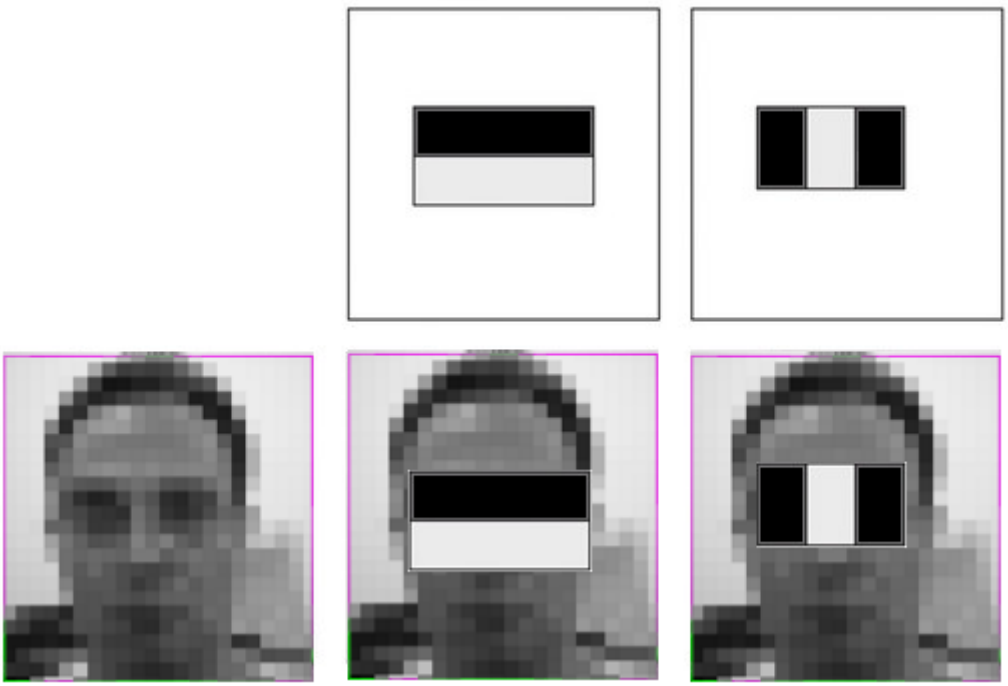

Figura 3. Primul clasificator la lucru pe o imagine 24 x 24

Astfel, cu toate că iniţial s-au străduit să implementeze un clasificator puternic prin combinarea a circa 200 de clasificatoare slabe, acest succes timpuriu i-a determinat ca în cele din urmă să construiască mai degrabă o cascadă de clasificatoare cu două clase fiecare, sub forma unui arbore de decizie degenerat, în locul unui singur clasificator uriaş. Această cascadă de filtre, denumită „cascadă atenţională" (Figura 4), reprezintă cea de-a treia inovaţie semnificativă introdusă de metoda Viola-Jones.

Fiecare subfereastră a imaginii originale este testată cu primul clasificator. Dacă trece de acesta, este testată cu al doilea. Dacă trece şi de acesta, este testată cu al treilea şi aşa mai departe. Dacă nu trece de un anumit nivel, subfereastra este rejectată ca posibilă faţă la nivelul respectiv. Numai dacă trece de toate filtrele din cascadă, atunci este clasificată ca fiind o faţă.

Ce este interesant, este că cel de-al doilea clasificator şi următorii nu (mai) sunt antrenaţi pe întregul set de învăţare, ci numai pe acele imagini care nu au fost rejectate de clasificatorii anteriori din lanţ. În acelaşi timp, subferestrele fals-pozitive scăpate de primele niveluri sunt oferite ca exemple de învăţare de non-feţe nivelurilor următoare.

Al doilea clasificator şi următorii sunt mai complecşi şi au mai multe caracteristici decât primul şi prin urmare sunt şi mai mari consumatori de timp computaţional. Este cu atât mai remarcabil faptul că există un astfel de clasificator simplu care să elimine atât de multe subferestre fără a necesita efectuarea calculelor necesare clasificatorilor mai complecşi. Pe de altă parte, este mult mai probabil ca imagini aleatoare să nu conţină efectiv o faţă, sau ca regiunile conţinând feţe în acestea să fie relativ puţine, iar faptul că majoritatea subferestrelor pot fi eliminate ca posibili candidaţi cu efort computaţional extrem de redus este un lucru benefic.

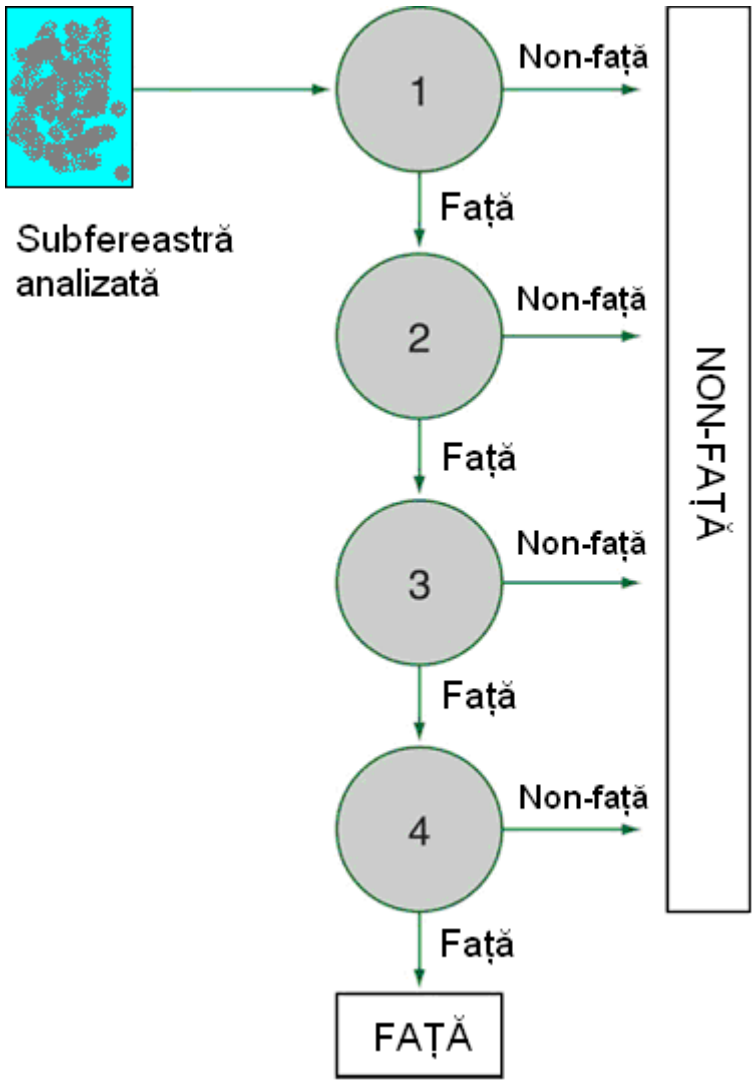

Figura 4. Cascada de clasificatori Viola-Jones

În cele din urmă, Viola şi Jones au utilizat 38 de niveluri pentru cascada de clasificatori [5], utilizându-se în primele 5 dintre acestea, respectiv, câte 2, 10, 25, 25 şi 50 de caracteristici, iar în total, pentru toate nivelurile, numărul acestora fiind de 6060. Numărul de caracteristici pentru fiecare nivel a fost stabilit empiric, prin încercare şi eroare, pentru primele niveluri. Criteriul a fost minimizarea fals-pozitivelor până sub un anumit prag pentru fiecare nivel, concomitent cu menținerea unei rate înalte a recunoaşterilor corecte (rată foarte scăzută a fals-negativelor). Au fost adăugate caracteristici fiecărui nivel, în etape, până când criteriile de performanță propuse la nivelul respectiv au fost atinse.

Clasificatoarele de pe primele niveluri sunt mai simple şi asigură eliminarea foarte rapidă a unui număr mare de subferestre de tip non-față. Astfel, primul clasificator din cascadă utilizează cele două caracteristici descrise mai sus şi elimină peste 50% dintre non-fețe, iar următorul, utilizând alte 10 caracteristici, elimină peste 80% dintre non-fețe.

Pentru învățare s-au utilizat 4.916 exemple de fețe descărcate aleator de pe Internet. decupate şi aliniate grosier (manual) şi scalate la 24 x 24 de pixeli. În procesul de învățare au fost utilizate ca exemple de fețe şi imaginile în oglindă ale acestora față de axa centrală verticală, deci un total de 9.832 de imagini. Exemplele de non-fețe au fost selectate ca subferestre din alte peste 9.500 de imagini aleatoare, de asemenea descărcate de pe internet şi verificate manual că nu conțin fețe. Numărul acestor subferestre este de circa 350.000.000, dar numai câte maximum 10.000 au fost utilizate pentru antrenarea fiecărui clasificator din cascadă. Mai trebuie precizat că este vorba de o învățare adaptivă. Dacă pentru primul nivel s-au utilizat subferestre din cele peste 9.500 de imagini

non-fețe, pentru nivelurile următoare s-au utilizat subferestrele de tip non-față obținute din fals-pozitivele date de nivelurile anterioare.

Întregul proces de învățare a durat mai multe săptămâni. Nu mai vorbim despre durata experimentărilor pentru alegerea caracteristicilor, a tipului de clasificator, pentru ajustarea ponderilor ș.a.m.d. Se poate concluziona că este vorba despre o metodă care a presupus o primă etapă extrem de laborioasă, din categoria celor în care învățarea este lentă, dar detecția este foarte rapidă.

În final, când se analizează o imagine curentă, detectorul de fețe scanează imaginea completă prin subferestre la mai multe scale şi cu mai multe poziționări pentru fiecare scală. Cercetătorii au stabilit că utilizarea unui factor de 1,25 de la o scalare a subferestrei la alta a produs cele mai bune rezultate. Totodată s-a mai constatat că nu este necesar să se testeze neapărat fiecare locație în parte. Chiar dacă se sar câțiva pixeli de fiecare dată la translatarea / glisarea subferestrei de analizat, rezultatele bune nu sunt afectate. Toate aceste observații empirice au permis o mai mare optimizare şi o îmbunătățire suplimentară a vitezei de calcul.

Ulterior, pentru a face mai bine față situațiilor în care figurile nu apar frontal în imagini, atât Jones şi Viola [6] cât şi alți cercetători, cum ar fi Lenhart şi Maydt [7] au continuat să aducă îmbunătățiri prin extinderea setului de caracteristici utilizate, introducând filtre diagonale [6] sau caracteristici rotite cu 45 de grade [7], dar au mai existat şi alte nenumărate contribuții care au adus completări şi îmbunătățiri punctuale metodei.

Mai amintim aici şi faptul că în biblioteca de funcții de vedere artificială (*Computer Vision*) dezvoltată şi menținută (inițial) de Intel, OpenCV, open source, disponibilă liber pentru mai multe platforme, există o implementare a metodei Viola-Jones ce poate fi direct utilizată prin includere a funcțiilor şi structurilor respective în diverse aplicații. Un exemplu este prezentat pe larg în [8].

4. Concluzii

În prezentul articol s-a încercat o introducere în domeniul detecției automate a fețelor, cu o prezentare succintă a principalelor tipuri de abordări şi o descriere ceva mai detaliată, dar totuşi la un nivel de înțelegere cât mai general, a unei metode de referință în domeniu, metoda Viola-Jones.

Detecția automată a fețelor în imagini digitale şi-a făcut loc în viața noastră de zi cu zi fiind inclusă ca funcționalitate în camerele foto digitale, aplicațiile de recunoaştere a fețelor incluse în sisteme de securitate, control automat al accesului, verificare a identității, în noile paradigme în ceea ce priveşte interfețele om-calculator evoluate şi clasificarea şi etichetarea datelor şi imaginilor pe Internet cu meta-informații şi acestea sunt numai câteva dintre cele mai la îndemână exemple.

Problema detectării figurilor umane în imagini nu este tocmai una trivială, datorită varietății uriaşe în care acestea pot apărea în percepția senzorului 2-D, atât din cauza trăsăturilor şi particularităților fizionomice, culorii, dimensiunilor, poziției, acoperirii parțiale, fundalului complex, dar mai cu seamă din cauza zonelor de lumină şi umbră determinate de poziția şi distribuția sursei / surselor de lumină.

Abordările şi soluțiile încercate de-a lungul timpului au trebuit să-şi propună să facă față tuturor situațiilor care pot fi întâlnite, indiferent de particularitățile subiectului şi în orice context de poziționare, dimensiune, scalare, încadrare, fundal, sau iluminare s-ar afla acesta. O altă necesitate pentru majoritatea, datorită specificului aplicațiilor, a fost să producă rezultate în timp real. Unele au reuşit mai bine, altele mai puțin bine. În evaluarea performanțelor sistemelor de detecție a fețelor se ține cont, pe lângă factorul viteză de calcul acolo unde aceasta este necesară, de două componente: rata fals-pozitivelor (confuziilor, raportări false ca fețe ale unor regiuni din imagini care nu conțin în realitate o față), precum şi rata fals-negativelor (rateurilor, regiuni ale imaginilor care în realitate conțineau o față, dar care nu au fost raportate ca atare), care trebuie să fie ambele cît mai mici, ideal chiar zero fiecare.

În general, metodele utilizate, indiferent de tipul acestora, sunt laborioase, presupun numeroase iterații, analize pe diferite criterii, scalări, filtrări, comparații, şi evident au solicitat eforturi şi timp pentru a fi puse la punct. Unele se bazează pe antrenări anterioare ale unor clasificatoare (în general

cu două clase: față şi non-față) de diferite tipuri (utilizând rețele neuronale, SVM etc.), în timp ce altele încearcă să se bazeze pe cunoştințe şi observații a priori codate programatic şi utilizate astfel la detecție, iar altele încearcă o mixare între aceste două tipuri de abordări în diferite etape de analiză.

O metodă de referință, care a marcat un punct de cotitură începând cu anul 2001 în evoluția sistemelor utilizând detecția fețelor, extrem de remarcabilă atât prin ingeniozitate cât şi prin performanțe şi viteză, este metoda Viola-Jones.

Spre deosebire de altele, metoda nu presupune o abordare piramidală prin scalarea imaginii inițiale cu diferiți factori şi totodată nu recurge la o analiză în sens clasic a valorilor intensităților pixelilor ce formează imaginea originală. Metoda recurge la analizarea unor caracteristici locale din imagini, utilizând aşa numite „caracteristici de tip Haar" (*Haar-like features*), prin analogie, dar fără a se confunda cu *wavelet*-uri Haar, şi care sunt reprezentate fiecare de câte 2, 3 sau 4 dreptunghiuri adiacente, cu tentă închisă şi respectiv deschisă, de formă şi dimensiuni identice între ele. Aceste caracteristici sunt scalate succesiv şi mutate prin subferestre de diferite dimensiuni, glisate peste imaginea originală şi pentru fiecare se calculează câte o valoare, dată de diferența între suma pixelilor din regiunile din imaginea originală acoperite de dreptunghiurile deschise minus suma pixelilor din regiunile acoperite de dreptunghiurile închise componente ale caracteristicii respective.

O altă inovație care a adus o considerabilă optimizare computațională, o reprezintă utilizarea unei „imagini integrale" calculate cu o singură trecere prin imaginea originală, în care fiecare punct capătă valoarea sumei tuturor pixelilor din stânga şi de deasupra pixelului corespondent din imaginea originală, inclusiv acesta. Suma pixelilor din orice dreptunghi definit pe imaginea originală se poate obține prin trei operații elementare (două scăderi şi o adunare) între cele patru valori din imaginea integrală corespunzătoare colțurilor dreptunghiului respectiv.

Pe baza unui prag ce se stabileşte empiric, se poate indica faptul că o caracteristică este prezentă sau nu într-o subfereastră analizată. În loc să construiască un unic clasificator complex, utilizând un clasificator de tip AdaBoost, Viola şi Jones au selectat cele mai discriminante caracteristici, pe care le-au combinat în mai multe niveluri, într-o cascadă de filtre, clasificatoare AdaBoost mai slabe fiecare („cascadă atențională" care reprezintă a treia inovație importantă introdusă de Viola şi Jones), în care primele niveluri, care includ mai puține caracteristici şi implică mai puține calcule, elimină masiv şi rapid subferestrele de tip non-față, urmând ca nivelurile următoare, mai complexe şi cu mai multe caracteristici, să analizeze mai profund subferestrele care au trecut de nivelurile anterioare dacă sunt fețe sau nu.

Pentru antrenare / învățare au fost utilizate aproape 5.000 de exemple de fețe şi 10.000 de exemple de non-fețe. Învățările au durat săptămâni.

Se poate vorbi despre o metodă în care învățarea este lentă, dar detecția este foarte rapidă. Rezultatul este că metoda Viola-Jones produce rezultate extrem de bune foarte rapid, suficient de repede pentru ca puterea de calcul limitată a camerelor digitale de clasă medie să poată face față cu succes.

O implementare a metodei Viola-Jones este disponibilă liber pentru mai multe platforme în biblioteca *open source* de funcții de vedere artificială (*Computer Vision*) a Intel, OpenCV.

BIBLIOGRAFIE

1. **YANG, M.-H.; KRIEGMAN, D. J.; AHUJA, N.:** Detecting Faces in Images: A Survey, in IEEE Trans. on PAMI, 24(1), pag. 34-58, 2002
2. **BUCKNALL, J. M.:** How to Find a Face, in PC Plus, Issue 296, July 18th 2010.
3. **GUPTA, R.; SAXENA, A. K.:** Survey of Advanced Face Detection Techniques, in Image Processing, International Journal of Computer Science and Management Research, Vol. 1, Issue 2, ISSN 2278-733X, September 2012, pp. 156-164.

4. **VIOLA, P.; JONES, M.:** Rapid Object Detection using a Boosted Cascade of Simple Features, in the Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2001.
5. **VIOLA, P.; JONES, M. J.:** Robust Real-Time Face Detection, in the International Journal of Computer Vision (IJCV) 57(2), pag. 137-154, Kluwer Academic Publishers, 2004.
6. **JONES, M.; VIOLA, P.:** Fast multi-view face detection, in Technical report, Mitsubishi Electric Research Laboratories, TR2003-96, 2003.
7. **LIENHART, R.; MAYDT, J.:** An extended set of Haar-like features for rapid object detection, in Proc. of ICIP, 2002.
8. **HEWITT, R.:** Seeing With OpenCV, Part 2: Finding Faces in Images, in SERVO Magazine, T & L Publications Inc., February 2007.
9. **FREUND, Y.; SCHAPIRE, R. E.:** A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, in Proceedings of EuroCOLT, 1995, pp. 23-37.