**UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI**

**Şcoala Doctorală de Electronică, Telecomunicaţii şi Tehnologia Informaţiei**
**Decizie nr.** 569 **din** 25-09-2020

# REZUMAT TEZĂ DE DOCTORAT


## Ing. Mihai Gabriel CONSTANTIN


---

# AUTOMATIC ANALYSIS OF THE VISUAL IMPACT OF MULTIMEDIA DATA

# ANALIZA AUTOMATĂ A IMPACTULUI VIZUAL AL DATELOR MULTIMEDIA

---

**COMISIA DE DOCTORAT**

| | |
|---|---|
| **Prof. Dr. Ing. Gheorghe BREZEANU**<br>Univ. Politehnica din Bucureşti | Preşedinte |
| **Prof. Dr. Ing. Bogdan IONESCU**<br>Univ. Politehnica din Bucureşti | Conducător de doctorat |
| **Prof. Dr. Ing. Martha LARSON**<br>Radbound University | Referent |
| **Dr. Ing. Claire-Hélène Demarty**<br>InterDigital | Referent |
| **Prof. Dr. Ing. Mihai CIUC**<br>Univ. Politehnica din Bucureşti | Referent |

**BUCUREŞTI 2021**

**Capitolul 1: Introducere** **3**
1.1 Domeniul tezei 3
1.2 Motivarea tezei 3
1.3 Conținutul tezei 4

**Capitolul 2: Aspecte teoretice** **4**
2.1 Taxonomie și definiții 4
2.2 Înțelegerea umană a proprietăților subiective ale datelor multimedia 5
2.3 Baze de date și studii pe utilizatori 5
2.4 Metode computaţionale 6
2.4.1 Grad de interes 6
2.4.2 Estetică 6
2.4.3 Memorabilitate 6
2.4.4 Violenţă 7
2.4.5 Valoare afectivă și emoţii 7

**Capitolul 3: Contribuții personale** **7**
3.1 Baze de date și evaluare 7
3.1.1 Predicția interesului 7
3.1.2 Predicția violenței 9
3.1.3 Predicţia memorabilităţii 9
3.1.4 Recomandare de conţinut 9
3.2 Predicția gradului de interes 10
3.2.1 Introducere 10
3.2.2 Sisteme de învățare bazate pe SVM 10
3.2.3 Descriptori estetici și sisteme de fuziune târzie 11
3.3 Predicția scenelor violente 13
3.3.1 Introducere 13
3.3.3 Sisteme de învățare adâncă temporală 13
3.4 Predicţia memorabilităţii 14
3.4.1 Introducere 14
3.4.2 Sisteme de învățare adâncă bazate pe acțiune 15
3.5 Fuziune târzie cu sisteme de asamblare adâncă 16
3.5.1 Introducere 16
3.5.2 Motivaţie 16
3.5.3 Lucrări anterioare 17
3.5.4 Metoda propusă 17
3.5.5 Setare experimentală 20
3.5.6 Rezultate experimentale 21

**Capitolul 4 - Concluzii generale și perspective** **21**
4.1 Contribuții și publicații 21
4.2 Concluzii 25
4.3 Perspective de viitor 26

# Capitolul 1: Introducere

## 1.1 Domeniul tezei

Această teză acoperă mai multe aspecte teoretice și metode de ultimă generație ce privesc analiza automată a impactului vizual al datelor multimedia. În timp ce sarcinile tradiționale de viziune computerizată vizează probleme care au un grad de adevăr obiectiv, cu care majoritatea adnotatorilor ar fi de acord, direcțiile recente de cercetare tind să studieze concepte subiective, cum ar fi gradul de interes, estetică, violența etc. În aceste cazuri, valoarea de adevăr poate depinde de un mare set de factori legați de om, inclusiv, preferințele personale, mediul cultural și starea psihologică actuală.

## 1.2 Motivarea tezei

Această teză își propune să contribuie la înțelegerea unui set de concepte subiective, să descopere și să sublinieze câteva practici de succes pentru cercetarea unor astfel de concepte și să dezvolte metode automate care le pot prezice cu precizie. În timp ce colecția extinsă de concepte prezintă grade diferite de subiectivitate și, prin urmare, gradul de încredere inter și chiar intra-evaluator în ceea ce privește imaginea adnotată și eșantioanele video din seturile de date precizate poate varia semnificativ, interesul pentru metodele automate care rezolvă aceste probleme și prezic aceste concepte este în creștere, indiferent de dificultățile create de subiectivitatea inerentă a conceptelor. Există o cerere din ce în ce mai mare pentru aceste metode, în mare parte determinate de platformele de socializare, distribuție, publicitare și arhivare media, care ar beneficia prin crearea unor predictori automați. Sisteme de recomandare bazate pe aceste concepte, filtre automate și alte funcționalități ar fie imposibil de implementat fără ajutorul viziunii computerizate, al învățării automate și al inteligenței artificiale.

Interesul și sprijinul din partea industriei pentru acest tip de cercetare este demonstrat până acum prin intermediul mai multor aplicații online, care reprezintă părți și module ale unor platforme mai mari, precum și organizarea de competiții de benchmarking care ajută atât comunitatea cercetătorilor, cât și industria. Flickr social interestingness application[1] și Google Photos summary creation[2] reprezintă unele dintre cele mai populare aplicații industriale, în timp ce InterDigital[3] a creat mai multe seturi de date și competiții de benchmarking privind interesul, violența și memorabilitatea.

[1] https://www.flickr.com/
[2] https://www.google.com/photos/about/
[3] https://www.interdigital.com/datasets/

## 1.3 Conținutul tezei

Restul acestei teze este împărțit în 3 capitole. Primul prezintă stadiul actual al tehnologiei în ceea ce privește taxonomiile, studiile psihologice, seturile de date, studiile pe utilizatori și abordările automate de predicție dezvoltate de cercetători din diferite domenii care se ocupă de problema definirii și predicției proprietăților subiective ale datelor multimedia. Al doilea capitol prezintă contribuțiile personale la acest domeniu, în ceea ce privește seturile de date și competiţiile la care am contribuit, metodele și modelele de calcul originale pentru predicția unora dintre aceste concepte, precum și o colecție de metode de fuziune târzie bazate pe rețele adânci, folosind o selecție largă de inductori de intrare mai slabi ca performanţa. Teza se încheie cu câteva concluzii generale și perspective pentru lucrările viitoare, precum și un rezumat al lucrărilor mele și contribuția la aceste lucrări.

# Capitolul 2: Aspecte teoretice

În domeniul internetului și big data, utilizatorii sunt bombardați în mod constant cu cantități mari de date multimedia, uneori devenind chiar și creatori de conținut, prin colecții de fotografii personale, postări pe rețelele sociale sau vloguri. Este într-adevăr dificil să se țină evidența tuturor acestor informații. Cercetătorii au arătat că această alimentare constantă cu informații, atât vizuale, cât și de altă natură, poate reduce semnificativ atenția vizitatorilor [1]. Astfel, apare nevoia de sisteme care pot procesa automat date și să filtreze sau să creeze liste de sugestii în funcție de preferințele utilizatorilor. Una dintre cele mai grele provocări cu care se confruntă aceste sisteme este reprezentată de definițiile acestor concepte, considerând că, spre deosebire de sarcinile mai tangibile, cum ar fi detectarea unui obiect dintr-o imagine, de cele mai multe ori, este greu pentru subiecții umani să cadă de acord asupra a ceea ce este interesant, estetic, violent și așa mai departe. Natura subiectivă a acestor concepte face din predicția și clasificarea lor una dintre sarcinile cele mai provocatoare din viziunea computerizată astăzi.

Acest capitol va prezenta o revizuire și o analiză a literaturii axate pe concepte care vor fi utilizate de-a lungul tezei, și anume *gradul de interes*, *estetica*, *memorabilitatea, violența* și *valoarea afectivă și emoțiile*.

## 2.1 Taxonomie și definiții

Primul concept analizat în această teză, *interesul* a fost definit ca un factor primar pentru motivație și un stimulent comportamental important pentru oameni [2, 3]. Hidi și Anderson [4] propun că atractivitatea unei activități poate fi mai importantă în generarea interesului decât preferințele personale. *Valoarea estetică* este definită ca o ramură a filozofiei care studiază atracția și frumusețea compozițiilor [5]. Din punct de vedere vizual, *memorabilitatea* este definită ca o proprietate intrinsecă a specimenelor

vizuale care măsoară probabilitatea ca subiecții să-și amintească imaginile și videoclipurile care le sunt prezentate. Mulți autori folosesc memorabilitatea pe termen scurt și lung [6] în definirea perioadei în care subiectul poate păstra informațiile memorate. În ceea ce privește *violența*, unii autori [7] propun atât o definiție subiectivă (mostre vizuale „pe care nu le-ar lăsa să le vadă un copil de opt ani, deoarece conțin violență fizică"), cât și o definiție obiectivă („violență fizică sau accident care are ca rezultat rănirea oamenilor sau durere") pentru violență. În cele din urmă, *valoarea afectivă și emoțiile* sunt definite ca fiind capacitatea informațiilor media de a induce anumite răspunsuri emoționale în subiecți [8]. Aceste emoții sunt descrise în principal în două moduri: fie într-un spațiu matematic 2D sau 3D cu excitare, valență și dominanță ca trăsături principale [9], fie într-un spațiu categoric, conținând emoții precum furia, frica, bucuria, surpriza etc. [10].

## 2.2 Înțelegerea umană a proprietăților subiective ale datelor multimedia

Studierea modului în care oamenii percep și interacționează cu datele multimedia este vitală pentru acest domeniu, deoarece creează un fundament puternic care ajută oamenii de știință din domeniul viziunii computerizate, oferind un set de principii cu ajutorul cărora pot fi dezvoltate metode de predicție automate.

*Gradul de interes.* Berlyne [11] și Silvia [3] identifică mai mulți factori care influențează interesul general, inclusiv noutatea, complexitatea, incertitudinea și conflictul. Cu toate acestea, după cum se arată în [12] aceste relații pot fi destul de complexe și neliniare. Dintr-o perspectivă evolutivă, Izard și Ackerman [13] concluzionează că interesul permite oamenilor să exploreze, să învețe și să interacționeze cu mediul lor. Reber et al. [14] propune că „bunătatea formei, simetria și contrastul figură-temelie" sunt calități pe care un element trebuie să le aibă pentru a fi considerate plăcute din punct de vedere *estetic*. Autorii au propus un set de „reguli ale fotografiei" care trebuie luate în considerare în analiza calității estetice a probelor vizuale [15]. Așa cum arată majoritatea studiilor din domeniul *memorabilităţii* [16], mintea umană are o capacitate impresionantă și poate neașteptată de a-și aminti datele vizuale. Arendt [17] studiază *violența* dintr-o perspectivă modernă, trecând prin unii dintre factorii săi posibili, cum ar fi „putere, forță, forță și autoritate". În același timp, Galtung [18] încearcă să o studieze dintr-o perspectivă culturală, observând diferența interculturală a percepției violenței. În final, *emoțiile* au fost studiate din multe perspective, variind de la teoria culorii [19] la o perspectivă educațională [20].

## 2.3 Baze de date și studii pe utilizatori

Colectarea unui set de date adecvat reprezintă unul dintre cele mai critice aspecte preliminare ale creării sistemelor automate pentru a prezice astfel de proprietăți subiective. Deși seturile de date sunt esențiale în general pentru sarcinile de învățare automată, în acest caz particular, trebuie luate în considerare unele aspecte

suplimentare, cum ar fi diferența de opinie între adnotatori, având în vedere subiectivitatea inerentă a datelor multimedia analizate.

Unele dintre cele mai importante seturi de date sunt reprezentate de lucrări care analizează mai multe concepte. Un exemplu din această categorie este setul de date visInterest [21] care are interesul drept concept principal, dar include și concepte care sunt în teorie pot influența interesul, cum ar fi potențialul de înțelegere, complexitatea și excitarea.

Alte seturi sunt construite în jurul ideii unei competiții comune de evaluare și oferă nu numai date și adnotări, ci și un set de descriptori, metrici, diviziuni de date, creând un mediu în care analiza performanței metodelor poate fi efectuată corect. Astfel de seturi de date sunt construite în jurul gradului de interes [22], memorabilitate [6] și violență [23].

# 2.4 Metode computaţionale

## 2.4.1 Grad de interes

În timp ce multe abordări computaționale au fost testate pentru a prezice interesul în media, până în prezent rețelele neuronale adanci nu au obținut performanțe optime. În timp ce unii autori încearcă să utilizeze concepte conexe pentru prezicerea gradului de interes, cum ar fi noutatea și estetica [24], de obicei prin utilizarea descriptorilor tradiționali, alții folosesc descriptorii direct pentru prezicerea interesului [25]. Competiția MediaEval Predicting Media Interestingness [22, 26] a dat posibilitatea de a testa mai multe sisteme în aceeași configurație în ceea ce privește setul de date, divizarea datelor, testare și metrica.

## 2.4.2 Estetică

Mai multe lucrări își bazează abordarea pe studii umane anterioare privind estetica, compoziția și regulile generale de fotografie. Unele lucrări esențiale aici includ [15, 27, 28]. Acești autori au proiectat un set cuprinzător de descriptori vizuali tradiționali care se bazează pe percepția umană și care sunt capabili să codifice cu exactitate unele dintre aceste principii, cum ar fi adâncimea de câmp, regula treimilor și combinațiile de nuanțe „plăcute”, proporțiile obiectelor , etc.

## 2.4.3 Memorabilitate

Metodele inițiale pentru predicția memorabilității [29] îmbină studiile umane cu metodele de viziune computerizată pentru clasificarea imaginilor, folosind concluziile trase din prima în proiectarea acesteia din urmă. Abordările mai moderne folosesc pe deplin puterea rețelelor neuronale adanci. De exemplu, mecanismele de atenție vizuală și straturile LSTM [30] sunt implementate într-o arhitectură convoluțională bazată pe ResNet de către Fajtl et al. [31].

### 2.4.4 Violenţă

Așa cum era de așteptat, majoritatea abordărilor pentru prezicerea acestui concept se bazează pe evaluare video în loc să utilizeze predicția unei singure imagini, deoarece violența este un concept inerent temporal. Unele exemple includ utilizarea descriptorilor de mișcare tradiționali [32], fluxului vectorial [33] sau abordări bazate pe LSTM [34].

### 2.4.5 Valoare afectivă și emoţii

Un larg segment din literatură este dedicat predicției conținutului emoțional. Zhao et al. [35] explorează un set de descriptori de nivel înalt bazați pe armonie și proporțiile dintr-o imagine, legând atracția estetică a imaginilor de emoțiile pe care le transmit. Descriptori specializati "sentiment" [36] și descriptori bazați pe excitare [37] sunt de asemenea utilizate pentru indicarea conținutului emoțional.

# Capitolul 3: Contribuții personale

## 3.1 Baze de date și evaluare

Acest capitol prezintă contribuțiile mele la crearea mai multor seturi de date disponibile publicului, inclusiv: (i) Interestingness10k, [38], conceput pentru predicția interesului în imagini și videoclipuri; (ii) VSD96 [39], un set de date video pentru detectarea scenelor violente; (iii) MediaEval 2019 Predicting Media Memorability [6] un set de date compus din videoclipuri scurte care sunt adnotate cu valori de memorabilitate pe termen scurt și lung; și în cele din urmă (iv) MMTF-14k [40], un set de date pentru recomandare de filme.

### 3.1.1 Predicția interesului

Interestingness10k [38], disponibil public[4], este un set de date și un cadru comun de evaluare, conceput pentru predicția gradului de interes al imaginilor și a videoclipurilor, validat și testat în cadrul competițiilor MediaEval 2016 și 2017 Predicting Media Interestingness. Contribuțiile mele principale la acest set de date sunt reprezentate de: (i) analiza performanței generale a sistemelor din cadrul MediaEval; (ii) analiza influenței descriptorilor asupra modelelor de predicție utilizate în timpul competiției MediaEval; (iii) analiza capacităților de generalizare a modelelor de predicție; (iv) crearea unui set de recomandări cu privire la performanța sistemelor; (v) participarea la procesul de adnotare. Setul de date este compus din mostre de imagini și videoclipuri extrase din filme asemănătoare cu Hollywood, licențiate Creative Commons[5], împărțite în 7396 mostre în setul de dezvoltare și 2192 mostre în setul de testare, în cea mai nouă versiune a setului de date.

[4] https://www.interdigital.com/data_sets/interestingness-dataset
[5] https://creativecommons.org

Pentru *analiza generală a performanței*, am adunat toate sistemele participante din competițiile MediaEval și am analizat tendințele și îmbunătățirile. Cea mai importantă observație în acest caz este că performanța sistemelor s-a îmbunătățit între cele două ediții ale competiției, cu 25,75% pentru procesarea de imagini și 22,75% pentru procesarea video. De asemenea, este interesant de remarcat faptul că, deși performanța adnotatorilor umani este mai bună decât performanța sistemelor de predicție automată, nici oamenii nu ating niciodată o performanță aproape perfectă, rezultatele lor fiind sub MAP = 0,7.

*Analiza la nivel de descriptori* arată că șase tipuri principale de descriptori sunt utilizaţi de participanți: vizual, audio, mișcare, bazat pe învățarea adâncă, conceptual și textual. Multe sisteme utilizează mai multe tipuri de caracteristici în diferite scheme de fuziune, creând 18 combinații ale acestor descriptori. În medie, pentru imagini, descriptorii bazați pe învățare adâncă au o performanță mai bună (MAP = 0,2297), în timp ce pentru video descriptorii vizuali tradiționali au o performanță mai bună (MAP = 0,1798).

Analiza *capabilităților de generalizare* arată câteva concluzii interesante cu privire la sistemele participante. De exemplu, pentru imagini, sistemele de învățare adâncă care conțin o etapă de pre-antrenament chiar și pe un set de date necorelat prezintă performanțe mai bune decât sistemele care nu folosesc pre-antrenament. De asemenea, există o corelație a rezultatelor (calculate prin Corelaţia Pearson = 0,546) între performanțele sistemelor de predicție a imaginilor și a videoclipurilor similare, indicând faptul că adaptarea predictorilor de imagine la videoclipuri poate reprezenta un prim punct de plecare. În cele din urmă, performanța sistemului pentru videoclipurile mai lungi a fost superioară performanței pentru videoclipurile scurte (MAP@10 = 0,751 vs. 0,0562), indicând faptul că videoclipurile mai lungi creează o separare mai mare între cele două clase.

În cele din urmă, propunem un set de recomandări cu privire la performanța sistemului, care include următoarele idei::

- descriptorii adanci (pentru imagini) și descriptorii vizuali tradiționali (pentru videoclipuri) funcționează mai bine decât alte tipuri de descriptori;
- sistemele de fuziune târzie reprezintă un avantaj evident în comparație cu sistemele care utilizează fuziune timpurie sau fără fuziune, această observație fiind susținută și de modelul nostru de asamblare bazat pe DNN;
- sistemele care utilizează mai multe tipuri de clasificator sau regresor tind să depășească sistemele cu un singur clasificator;
- abordările DNN mai moderne, cum ar fi GSM-InceptionV3 [41], pot avea performanțe bune, dar nu depășesc abordările specifice domeniului;
- mărirea numărului de eșantioane are un efect pozitiv asupra performanței sistemului, așa cum se arată în [42];
- performanța sistemului poate fi îmbunătățită prin antrenarea în prealabil cu date externe [43].

### 3.1.2 Predicția violenței

Setul de date VSD96 [39] este un set de date disponibil public[6] [7] și un cadru comun de evaluare conceput pentru detectarea scenelor violente în filmele Hollywood și YouTube. Versiunile acestui set de date au fost utilizate în cadrul competiției de detectare a scenelor violente MediaEval 2011-2015. Principalele contribuții personale la acest set de date sunt: (i) o analiză generală a sistemelor care utilizează acest set de date și (ii) o analiză a tipurilor de caracteristici utilizate pentru predicția violenței.

*Analiza generală a sistemelor* arată că, în general, performanța sistemelor participante s-a îmbunătățit, ajungând la un MAP de 0.51 utilizând definiția obiectivă a violenței. De asemenea, sunt încurajatoare rezultatele bune înregistrate pe partea de generalizare YouTube a setului de date, dat fiind că sistemele nu au fost antrenate cu acel tip de date, arătând astfel o bună codare a conceptului general de violență.

*Analiza descriptorilor* utilizaţi arată că participanții au folosit în principal patru tipuri de descriptori: vizual, audio, conceptuali și învățare adâncă. În timp ce primele trei sunt utilizate pe parcursul tuturor edițiilor competițiilor, învățarea adâncă devine populară în 2014 și mai ales în ediția din 2015 a competiției. În general, 12 combinații ale acestor tipuri sunt utilizate de participanți. Mai mult, în ceea ce privește sistemele multimodale, se evidențiază patru categorii, obținând rezultate de top în anumite sarcini: (i) vizual și audio, (ii) audio și conceptual, (iii) vizual, audio și conceptual și (iv) vizual, audio și învățare adâncă. În cele din urmă, sistemele de fuziune târzie obțin o performanță MAP mai bună decât sistemele de fuziune timpurie sau unimodale.

### 3.1.3 Predicţia memorabilităţii

Baza de date MediaEval 2019 Predicting Media Memorability [6], este un set de date validat în cadrul ediției 2019 a competiției MediaEval. Această sarcină cere participanților să prezică cu precizie memorabilitatea pe termen scurt și lung a videoclipurilor. Pentru acest set de date, contribuția mea principală este conducerea echipei de organizare în timpul competiției MediaEval. Setul de date este format din 10000 de videoclipuri scurte fără sunet, împărțite între setul de dezvoltare (80% din date) și setul de testare (20% din date). În general, această ediție a competiției arată îmbunătățiri semnificative față de edițiile anterioare în ceea ce privește performanța sistemelor participante.

### 3.1.4 Recomandare de conținut

MMTF-14K [40] este o bază de date disponibilă public[8] care creează o colecție de date pentru sistemele de recomandare a filmelor de la Hollywood. În timp ce majoritatea sistemelor și bazelor de date de recomandare își bazează deciziile pe metadate, constând din evaluări ale utilizatorilor, genuri de filme și alți descriptori

[6] Data for 2011-2014 available at: https://www.interdigital.com/data_sets/violent-scenes-dataset
[7] Data for 2015 available at: http://liris-accede.ec-lyon.fr/
[8] https://zenodo.org/record/1225406.Xw830s8zaXw

asociați, acest set de date oferă în plus descriptori audio și vizuali care pot ajuta procesul de recomandare, creând un sistem de decizie multimodal. Contribuția mea principală la acest set de date este reprezentată de procesarea caracteristicilor vizuale bazate pe învățarea adânca și a descriptorilor estetici vizuali.

# 3.2 Predicția gradului de interes

## 3.2.1 Introducere

În acest capitol, prezentăm contribuțiile referitoare la prezicerea gradului de interes al datelor media. Propunem implementarea sistemelor de învățare bazate pe SVM care utilizează mai multe caracteristici vizuale [44], precum și sisteme de învățare bazate pe utilizarea descriptorilor estetici și fuziunea târzie [45, 46]. Contribuțiile principale constau în aplicarea unui set de descriptori vizuali tradiționali și a unui set de descriptori estetici în domeniul predicției interesului vizual și aplicarea schemelor de fuziune târzie pentru a îmbunătăți performanța finală a sistemului.

## 3.2.2 Sisteme de învățare bazate pe SVM

Această abordare constă din trei faze, așa cum se arată în Figura 3.2.1. Prima etapă implică extragerea unui set de descriptori vizuali tradiționali, urmată de o etapă care agregă descriptorii în diverse combinații și se termină cu o metodă de învățare bazată pe SVM.

Un set de șapte descriptori este extras pentru fiecare dintre imaginile și videoclipurile din setul de date: (i) histogramă de culoare în spațiul HSV, (ii) transformată DenseSIFT, (iii) LBP, (iv) HoG, (v) GIST, (vi) descriptori extrași din straturile fc7 și prob ale arhitecturii AlexNet [47], (vii) histograma de denumire a culorilor [48].

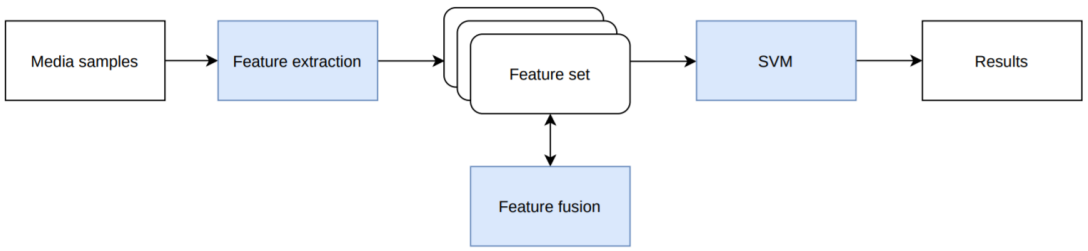


*Figura 3.2.1 Diagrama metodei bazate pe SVM propusă. Cele trei etape principale (extragerea descriptorilor, fuziunea lor și SVM) sunt evidențiate în albastru.*

Pentru predicția imaginilor, fiecare descriptor este extras individual și tratat ca un vector de valori în virgulă mobilă, în timp ce pentru predicția video, cadrele individuale sunt extrase și apoi agregate la nivel video prin media vectorilor cadrelor. Fuziunea se realizează prin concatenarea diferiților vectori de caracteristici.

Pentru a maximiza performanța sistemului, alegem un set larg de experimente și începem prin implementarea nucleelor polinomiale, RBF și liniare. Următorii parametri SVM sunt testați pentru nucleele polinomiale pentru a optimiza rezultatele:

- gradul ($d$) cu valori de 1, 2 si $3 \times k$, unde $k \in [1,\ldots, 10]$;
- coeficientul gamma ($\gamma$) cu valori de $2^k$, unde $k \in [1,\ldots, 6]$;

iar pentru nucleul RBF:

- cost ($c$)
- $\gamma$, ambele cu valori de $2^k$, unde $k \in [-4,\ldots, 8]$;.

*Setare experimentală*. Diversele combinații de descriptori și modele SVM sunt testate în contextul MediaEval 2016 Predicting Media Interestingness Task [26].

| | Sistem | MAP | P@5 | P@10 | P@20 | P@100 |
|---|---|---|---|---|---|---|
| imagini | ME top | 0.2336 | - | - | - | - |
| | ME avg | 0.2009 | - | - | - | - |
| | HSVHist+GIST | 0.1714 | 0.1077 | 0.1346 | 0.1423 | 0.0869 |
| | SIFT+GIST | 0.1398 | 0.0462 | 0.0808 | 0.1 | 0.0862 |
| video | ME top | 0.1815 | - | - | - | - |
| | SIFT+ANprob | 0.1629 | 0.1154 | 0.15 | 0.1192 | 0.0819 |
| | GIST+ANprob | 0.1574 | 0.0923 | 0.1269 | 0.1212 | 0.0812 |
| | ANfc7+HSVHist | 0.1572 | 0.1231 | 0.1 | 0.1077 | 0.0815 |
| | ME avg | 0.1572 | - | - | - | - |

*Tabel 3.2.1 Rezultatele sistemului pe setul de teste, care arată cele mai bune rezultate ale sistemelor trimise, comparativ cu sistemele medii și cele mai performante la concursul MediaEval.*

Tabelul 3.2.1 prezintă rezultatele pe setul de testare al sistemelor noastre cele mai performante. Având în vedere MAP, metrica oficială, obținem cele mai mari performanțe pentru sistemele dezvoltate cu o combinație HSVHist + GIST pentru imagini (MAP = 0.1714) și SIFT + ANProb pentru videoclipuri (MAP = 0.1629).

## 3.2.3 Descriptori estetici și sisteme de fuziune târzie

Având în vedere rezultatele anterioare [44] prezentate la MediaEval 2016, nevoia de a implementa metode care sunt adaptate la predicția interesului devine evidentă. Așa cum este prezentat în lucrarea noastră de sinteză a literaturii [38], estetica și interesul sunt des studiate împreună și corelate. Decidem să extragem un set de descriptori bazați pe estetică, dezvoltați în [15, 27, 28] și îi folosim în predicția interesului. Testăm această abordare pe datele din MediaEval 2016 [26] și 2017 [22] Predicting Media Interestingness Task, publicând această abordare în două lucrări [45, 46].

Figura 3.2.2 prezintă o diagramă a acestei abordări. La acest nivel, diferența dintre aceasta și metodele noastre anterioare este reprezentată de apariția unui al patrulea stadiu de fuziune târzie.

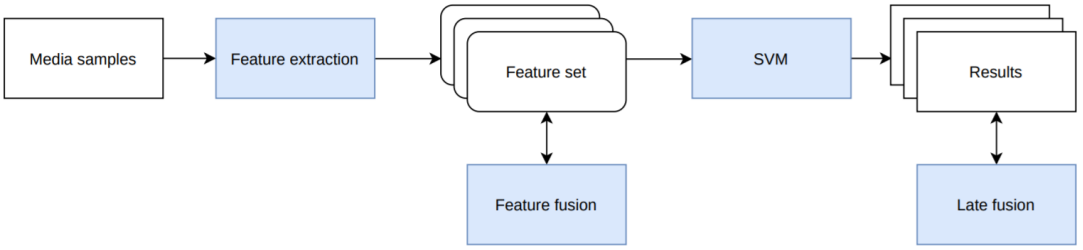


*Figura 3.2.2 Diagrama metodei bazate pe SVM propusă. Cele trei etape principale (extragerea descriptorilor, fuziunea lor, SVM și fuziunea târzie) sunt evidențiate în albastru.*

Abordarea noastră utilizează un set de clasificatori SVM cu nuclee polinomiale, RBF și liniare și augmentare prin fuziune timpurie și târzie.

În ceea ce privește descriptorii estetici, trei grupuri principale de caracteristici sunt utilizate în această lucrare, așa cum este descris în [49]: (i) caracteristici bazate pe culoare, (ii) caracteristici bazate pe textură și (iii) descriptori de obiect sau bazate pe segmentare. Unii din acești descriptori sunt inspirați de cercetările efectuate în domenii corelate, cum ar fi teoria culorilor, practicile fotografice și compoziția imaginii. Mai mult, SVM utilizează aceeași parametri ca în experimentele anterioare.

Metodele pe care le folosim în această lucrare sunt următoarele: (i) CombSum, (ii) CombMin, (iii) CombMax, (iv) CombMean. Prima dintre aceste metode constă în însumarea ieșirilor de predicție ale sistemelor inductoare, în timp ce CombMin și CombMax iau valoarea minimă și respectiv maximă a ieșirilor de predicție ale inductorului. Ultima metodă constă într-o medie ponderată a ieșirilor inductorului.

| Abordare | MAP | Descriere |
|---|---|---|
| Late fusion | 0.2485 | CombMax (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm) |
| Early fusion | 0.2363 | SatSegm + MassVarSegm + SkewSegm |
| ME top | 0.2336 | |
| Inducer | 0.2057 | aHSVWavelet or SatSegm |
| ME avg | 0.2009 | |

*Tabel 3.2.1 Rezultatele sistemului pe setul de teste, care arată cele mai bune rezultate ale sistemelor trimise, comparativ cu sistemele medii și cele mai performante la concursul MediaEval 2016.*

Tabelul 3.2.1 prezintă cele mai bune rezultate la competiția MediaEval 2016. Atât sistemele de fuziune timpurie, cât și cele târzii performează mai bine decat cel mai bun sistem din literatură. De asemenea, este interesant de remarcat faptul că inductorul de top a avut, de asemenea, performanțe peste rezultatele medii din MediaEval. Performanța maximă a fost obținută printr-un sistem CombMax de fuziune târzie.

| | Sistem | MAP testset | MAP@10 testset |
|---|---|---|---|
| imagini | ME top | 0.3075 | 0.1385 |
| | ME avg | 0.2402 | 0.0876 |
| | CombMean (aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm) | 0.1873 | 0.5555 |

| | | | |
|---|---|---|---|
| video | ME top | 0.2094 | 0.0827 |
| | CombMean(LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture) | 0.2028 | 0.0732 |
| | ME avg | 0.1845 | 0.0827 |

*Tabel 3.2.2 Rezultatele sistemului pe setul de teste, care arată cele mai bune rezultate ale sistemelor trimise, comparativ cu sistemele medii și cele mai performante la concursul MediaEval 2017.*

Tabelul 3.2.2 prezintă cele mai bune rezultate pentru competiția MediaEval 2017, unde din nou cel mai performant sistem este o abordare de fuziune târzie. În timp ce de această dată abordările propuse nu au depășit performanțele de top din competiția MediaEval, pentru performanța pe video a fost peste medie.

# 3.3 Predicția scenelor violente

## 3.3.1 Introducere

În această secțiune, prezentăm contribuția la predicția scenelor violente în filme și în videoclipuri de supraveghere extrase de pe YouTube. Această abordare utilizează o structură ConvLSTM [50] care procesează descriptorii vizuali creaţi prin procesarea diferențelor între cadre video cu o rețea VGG [51]. Experimentele cu această abordare sunt validate pe două seturi de date: setul de date MediaEval 2015 Violent Scene Detection [23] și setul de date VIF [33].

## 3.3.3 Sisteme de învățare adâncă temporală

Detectarea scenelor și evenimentelor violente este implică o componentă temporală inerentă; prin urmare, alegem să implementăm abordări de ultimă generație în ceea ce privește analiza secvențelor video. Algoritmul constă dintr-un model DNN temporal cu capacitatea de a aduna și recunoaște informații spațio-temporale din eșantioanele video. Sistemul nu utilizează direct cadrele video ca intrare pentru etapa de procesare, ci diferențele dintre cadrele video consecutive, așa cum este propus în [52], în ipoteza că rețelele vor fi instruite de la început cu o corelație de mișcare internă între hiperparametri. Diferențele de cadre sunt transmise după etapa inițială la un model VGG-19 [51], care va codifica un set de descriptori pentru fiecare pereche de diferențe de cadru. În faza finală, straturile ConvLSTM [50] vor procesa ieșirea rețelei VGG. Configurarea specială a stratului ConvLSTM pentru acest experiment este următoarea. Folosim 256 de filtre cu o dimensiune egală cu $3 \times 3$, obținând astfel o ieșire de 256 de caracteristici pentru fiecare segment video procesat. Videoclipurile sunt procesate cu o fereastră de cadre de dimensiuni variabile, echivalând cu aproximativ 1 secundă. Straturile finale sunt complet conectate cu o dimensiune de 512 neuroni fiecare și procesează ieșirea ConvLSTM pentru a obține o decizie finală. Această arhitectură de rețea este prezentată în Figura 3.3.1.

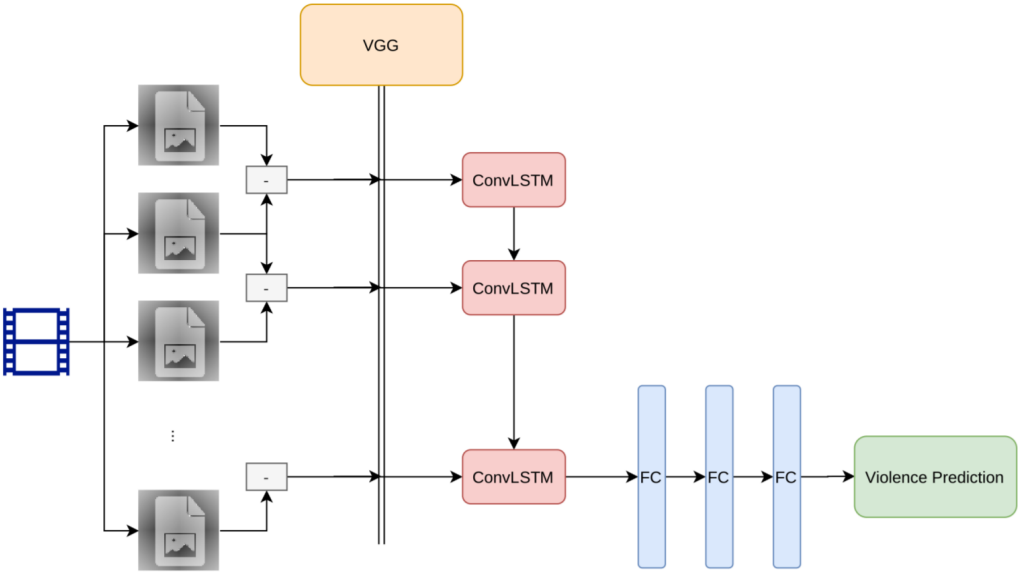


*Figura 3.3.1 Diagrama abordării propuse.*

Rezultatele experimentale sunt prezentate în Tabelul 3.3.1, unde sunt comparate cu rezultate din literatură pentru fiecare set de date. Rezultatele pentru această abordare sunt promițătoare, cu o valoare MAP maximă de 0,271 pentru setul de date VSD 2015, reprezentând o performanță mai mică în comparație cu rezultatul de top actual, care obține un MAP de 0,296, dar cu rezultate mai bune pe setul de date VIF, adică o acuratete de 0,89, comparativ cu rezultatele anterioare de top de 0,863.

| Metoda | Config. fereastră | VSD2015 (MAP) | VIF (Acc) |
|---|---|---|---|
| SOA | - | 0.296 | 0.863 |
| Sistem propus | 30 | 0.271 | 0.89 |

*Tabel 3.3.1 Rezultatele sistemului pe cele două seturi de date, comparate cu performanțele actuale de top ale fiecărui set de date.*

# 3.4 Predicţia memorabilităţii

## 3.4.1 Introducere

În acest capitol, prezentăm contribuțiile la predicția memorabilității. Lucrarea [53] propune implementarea sistemelor bazate pe estetică și recunoașterea acțiunilor în domeniul memorabilității și augmentarea rezultatelor prin implementarea unei etape finale de fuziune târzie. Contribuțiile mele la această lucrare sunt reprezentate de implementarea sistemelor bazate pe recunoașterea acțiunilor și implementarea schemelor de fuziune târzie. Abordările sunt testate un set de date public, publicat în cadrul MediaEval 2019 Predicting Media Memorability.

### 3.4.2 Sisteme de învățare adâncă bazate pe acțiune

În procesarea video, sistemele de recunoaștere a acțiunilor nou dezvoltate bazate pe rețele neuronale adânci reprezintă abordări de ultimă generație. Aceste rețele profită de straturi temporale, cum ar fi straturile LSTM [30], incluse în arhitecturile lor pentru a produce rezultate mai bune pe date temporale. Utilizarea unor astfel de rețele ar putea oferi rezultate bune pentru predicția memorabilității media prin codificarea precisă a caracteristicilor temporale asociate cu mostrele video.

Pentru această abordare, folosim mai multe modele DNN care sunt pre-antrenate pe estetică și recunoașterea de acțiuni. Pentru modelele bazate pe estetică, o arhitectură ResNet-101 [54] este antrenată pe datele de memorabilitate. În același timp, pentru DNN-urile de recunoaștere a acțiunii, rețelele TSN [55] și I3D [56] sunt utilizate pentru extragerea de descriptori și completate cu funcțiile C3D [57] furnizate de organizatori. Modelele de recunoaștere a acțiunii sunt trecute printr-un pas de reducere a dimensionalității, bazat pe PCA, iar antrenarea este realizată de un model SVR. Un ultim pas implică utilizarea schemelor de fuziune târzie. Schema acestei abordări este prezentată în Figura 3.4.1.

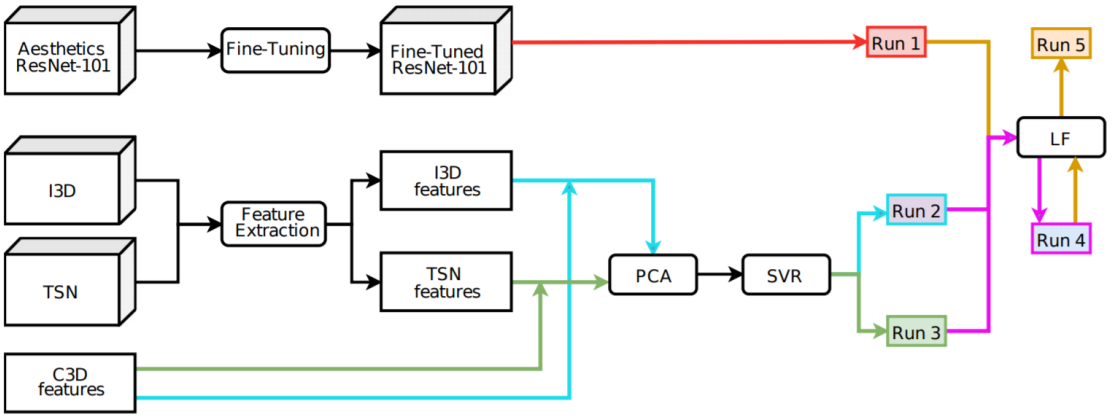


*Figura 3.4.1 Diagrama soluției propuse. Reprezentăm rețeaua bazată pe estetică (ResNet-101) și rețelele de recunoaștere a acțiunii (I3D, TSN și C3D), procesul de reglare fină sau extracție și procesul de învățare și etapa finală de fuziune târzie (LF). Sunt reprezentate și componentele celor cinci rulări individuale trimise la MediaEval Predicting Media Memorability (Run 1 - Run 5)*

Arhitectura bazată pe estetică este descrisă în Kang et al. [58]. Extragem stratul „Mixed_5” și îl folosim ca descriptor din modelul I3D, antrenat pe setul de date Kinetics [59], în timp ce stratul „Inception_5” 'este extras din modelul TSN, antrenat pe setul de date UCF101 [60]. Efectuăm teste preliminare în ceea ce privește I3D și TSN, dar și în ceea ce privește combinațiile lor de fuziune timpurie cu C3D. Aceste teste preliminare favorizează combinațiile timpurii de fuziune. În cele din urmă, un model SVR este utilizat pentru a antrena aceste descriptori într-o împărțire randomizată de 4 ori a datelor. Reglăm parametrii acestui model SVR folosind un nucleu RBF cu parametrii C și gamma luând valori de $10^k$, unde $k \in [-4,\dots 4]$. În cele din urmă, cele trei scheme de fuziune târzie pe care le folosim sunt CombMax, CombMin și CombMean.

| | | devset | | testset | |
|---|---|---|---|---|---|
| Run | System | short | long | short | long |
| | ME top | - | - | 0.528 | 0.277 |
| r5 | LF Aesthetic + Action (r1 + r2) | 0.494 | 0.265 | 0.477 | 0.232 |
| r2 | Action (TSN + I3D) | 0.473 | 0.259 | 0.45 | 0.228 |
| | ME avg | - | - | 0.448 | 0.206 |
| r4 | LF Action (r2 + r3) | 0.466 | 0.2 | 0.439 | 0.218 |
| r1 | Aesthetic | 0.448 | 0.23 | 0.401 | 0.203 |
| r3 | Action (C3D + I3D) | 0.433 | 0.204 | 0.386 | 0.184 |

*Table 3.4.1 Rezultatele sistemului pe setul de date Predicting Media Memorability, comparativ cu rezultatele de top.*

Rezultatele finale pe setul de testare, prezentate în tabelul 3.4.1, arată că cel mai performant sistem folosește o combinație târzie de fuziune a sistemelor estetice de predicție a rețelei și recunoașterea acțiunii. Două dintre modelele noastre au rezultate superioare rezultatelor medii MediaEval, și anume fuziunea timpurie a descriptorilor de acțiune TSN și I3D și abordarea de fuziune târzie care îmbină acțiunea și estetica. Pentru cel din urmă, cele mai bune rezultate sunt ρ= 0.477 pentru memorabilitatea pe termen scurt și ρ = 0.232 pentru cea pe termen lung.

# 3.5 Fuziune târzie cu sisteme de asamblare adâncă

## 3.5.1 Introducere

În acest capitol, prezentăm contribuțiile la crearea sistemelor de fuziune tarzie bazate pe rețele DNN. Lucrările [61, 62] și [38][9] propun crearea de sisteme de ansamblu care utilizează DNN-uri ca principalul motor de asamblare. Din câte știm, acest tip de abordare reprezintă o noutate în domeniul fuziunii informaționale, unde până acum DNN-urile au fost utilizate doar ca inductori pentru sistemele tradiționale de fuziune. Contribuția mea la această lucrare este reprezentată de (i) crearea a două noi scheme de transformare a intrării 2-D și 3-D care permit utilizarea straturilor neuronale adânci multidimensionale, (ii) implementarea straturilor convoluționale în sistemele de asamblare, (iii) ) și crearea unui nou strat DNN, special conceput pentru sistemele de fuziune, denumit stratul Cross-Space-Fusion. Sistemele propuse sunt testate pe mai multe seturi de date disponibile publicului, publicate ca parte a mai multor sarcini MediaEval, folosind ca inductori sistemele care au participat la sarcinile lor respective, după cum ne-au furnizat organizatorii sarcinilor.

## 3.5.2 Motivaţie

Așa cum s-a prezentat în unele dintre capitolele anterioare, sistemele de asamblare sau de fuziune târzie pot crește semnificativ performanța algoritmilor inductori pentru concepte subiective, cum ar fi interesul vizual și predicția memorabilității.

[9] Lucrare încă în stadiul de propunere

Descoperirile noastre în acest domeniu sunt susținute de alte lucrări, în care ansamblurile au reușit să obțină rezultate de top. Exemple în acest sens ar include interesul [63], memorabilitatea [64] și analiza conținutului emoțional [65], dar și domenii care nu tratează astfel de concepte subiective, exemple aici incluzând clasificarea acțiunilor umane în videoclipuri [41] .

### 3.5.3 Lucrări anterioare

Până în prezent, sistemele de fuziune târzie au folosit un set de metode tradiționale pentru asamblarea inductorilor. Unele exemple sunt deja prezentate în această teză, în principal metode statistice precum CombMin, CombMax, CombMean etc. Alte metode populare din literatură includ metode de boosting precum AdaBoost [66] și Gradient Boosting [67].

### 3.5.4 Metoda propusă

Abordarea DeepFusion propusă utilizează mai multe tipuri de straturi DNN care iau ca intrare setul de ieșiri inductoare și produc un nou set de ieșiri asamblate, în conformitate cu rezultatele pozitive și negative pe care rețeaua a reușit să învețe în timpul procesului de învățare. Astfel, propunem să începem cu crearea unui sistem de asamblare profundă de bază, o combinație de straturi dense de dimensiuni variabile. Această linie de bază va fi apoi augmentată prin adăugarea de straturi convoluționale și, în cele din urmă, prin adăugarea stratului Cross-Space-Fusion (CSF). În timp ce rețelele dense folosesc o intrare unidimensională pentru fiecare eșantion de imagine și video, straturile convoluționale și CSF utilizează intrări bidimensionale sau tridimensionale. Scopul acestor straturi este similar cu cel al convoluțiilor din procesarea imaginilor: vom încerca să descoperim și să învățăm corelații spațiale între valorile de intrare care sunt grupate spațial. Cu toate acestea, astfel de informații sunt imposibil de extras din vectorul 1-D de intrări care corespunde fiecărui eșantion, creat de ieșirile inductorilor individuali. Prin urmare, creăm un set de scheme de transformare a intrărilor care ne permit să construim structuri de intrare 2D și 3D, bazate pe gradul de similitudine dintre inductori individuali, făcând astfel posibilă implementarea straturilor convoluționale și CSF.

*Rețele dense.*
Straturile dense sunt cunoscute pentru că clasificarea datelor de intrare în categorii de ieșire cu precizie, reprezentând astfel o parte integrantă a tuturor abordărilor DNN. Având în vedere natura lor agnostică față de intrare, construirea unei rețele inițiale de bază care să integreze mai multe straturi dense ar reprezenta un punct de plecare bun în crearea rețelei. O reprezentare a unei arhitecturi dense este prezentată în Figura 3.5.1. Alegem să modificăm un set de parametri ai acestor rețele pentru a optimiza performanța. Se aleg următorii parametri: (i) numărul de straturi, (ii) numărul de neuroni pe strat și (iii) prezența sau absența straturilor de normalizare. Modificăm valorile acestor parametri până când se obțin cele mai bune rezultate pentru seturile de date alese.

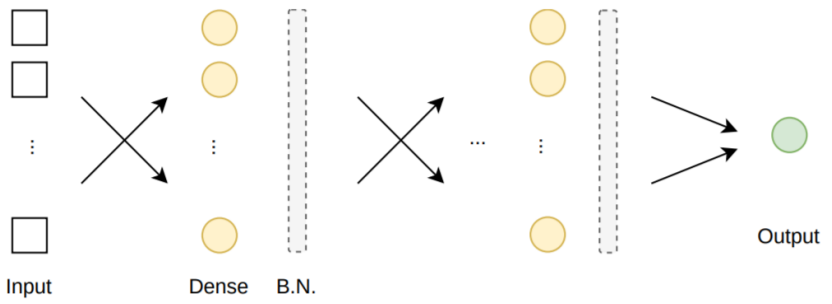


*Figura 3.5.1 Arhitectura de rețea densă DeepFusion (DF-Dense): număr variabil de straturi, numărul de neuroni pe strat și prezența sau absența straturilor de normalizare în lot (BN).*

*Decorarea intrărilor*

Pre-procesăm datele de intrare și să decorăm fiecare element cu scoruri de ieșire și date de la cei mai similari inductori pentru a genera informații spațiale. Având o imagine sau un eșantion video $s_i$, $i \in [1, M]$, fiecare dintre cei N algoritmi inductori va produce un set de ieșiri, $Y_i$, și, așa cum am menționat anterior, acest tip de intrare nu are o corelație spațială intrinsecă asociată cu aceasta. În primul pas al tehnicii de pre-procesare a intrării, analizăm corelația dintre inductorii individuali $f_i$, $i \in [1, N]$. Această corelație poate fi determinată prin orice metodă standard, cum ar fi scorul de corelație Pearson. Pentru a asigura un proces de învățare optimizat, vom folosi aceeași metodă ca cea utilizată drept metrică oficială.

După cum am menționat anterior, luăm în considerare atât o reprezentare 2D, cât și o reprezentare 3D a spațiului de intrare decorat. Pentru reprezentarea 2D, denumită *tr2D*, această schemă de decorare a intrării va fi utilizată pentru decorarea intrării pentru procesarea convoluțională. Pe de altă parte, cele două ecuații atribuite *tr3D* descriu reprezentarea 3D, fiecare dintre cele două matrice fiind stocate la indici diferiți în a 3-a dimensiune, creând o structură utilizată de stratul CSF.

$$tr2D_{i,j} = \begin{bmatrix} c_{1,i,j} & r_{1,i,j} & c_{2,i,j} \\ r_{4,i,j} & s_{i,j} & r_{2,i,j} \\ c_{4,i,j} & r_{3,i,j} & c_{3,i,j} \end{bmatrix},$$

$$tr3Dc_{i,j} = \begin{bmatrix} c_{1,i,j} & c_{2,i,j} & c_{3,i,j} \\ c_{8,i,j} & s_{i,j} & c_{4,i,j} \\ c_{7,i,j} & c_{6,i,j} & c_{5,i,j} \end{bmatrix}, tr3Dr_{i,j} = \begin{bmatrix} r_{1,i,j} & r_{2,i,j} & r_{3,i,j} \\ r_{8,i,j} & 1 & r_{4,i,j} \\ r_{7,i,j} & r_{6,i,j} & r_{5,i,j} \end{bmatrix}$$

Fiecare element $s_{i,j}$, reprezintă ieșirea de predicție produsă de inductorul $i$ pentru un eșantion $j$ de intrare în modelul propus, și este decorat cu ieșiri de la sisteme similare, $c_{1,i,j}$ reprezentând cel mai similar sistem, $c_{2,i,j}$ reprezentând al doilea cel mai

similar sistem și așa mai departe. Pentru valorile *r* introducem scorurile de corelație pentru cel mai similar sistem ($r_{1,i,j}$), al doilea cel mai similar ($r_{2,i,j}$) și așa mai departe, cu valoarea 1 în centru, care corespunde elementului $s_{i,j}$.

*Rețele dense cu straturi convoluționale*

O prezentare generală a arhitecturii convoluționale este prezentată în Figura 3.5.2. După procesarea intrării și transformarea acesteia într-o formă tr2D, această intrare este alimentată într-un strat convoluțional. Dată fiind acoperirea $3 \times 3$din fiecare element al intrării originale, alegem să folosim și un filtru $3 \times 3$ în arhitectura propusă, obținând astfel 10 parametri antrenabili în acest strat. Folosim un parametru de pas de 3, asigurându-ne că fiecare filtru convoluțional procesează doar sisteme similare. Această structură este urmată de un strat de pooling care va aduce ieșirea convoluției la forma inițială de intrare 1D. De asemenea, testăm 1, 5 și 10 filtre pe convoluție, permițând rețelei să efectueze o analiză mai extinsă a asemănărilor.

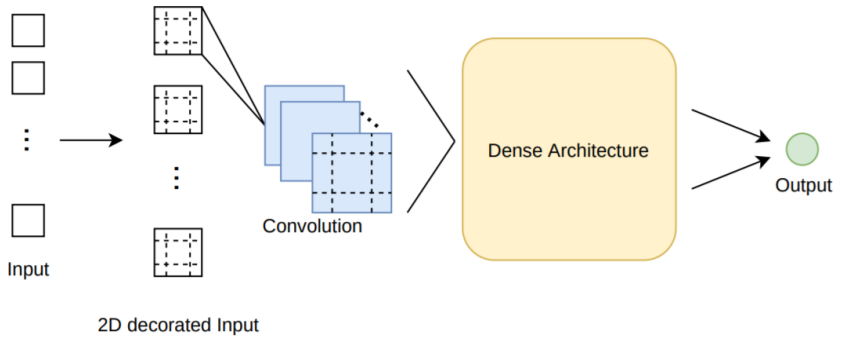


*Figura 3.5.2 Arhitectura de rețea convoluțională DeepFusion (DF-Conv). Aici sunt reprezentate stadiul de procesare a intrărilor, filtrele convoluționale și arhitectura densă finală.*

*Rețele dense cu straturi Cross-Space-Fusion*

În cele din urmă, introducem stratul Cross-Space-Fusion (CSF), al cărui design general este prezentat în Figura 3.5.3. Acest strat utilizeaza matricea 3D *tr3D* și, pentru fiecare grup de centroizi (*tr3Dc*, *tr3Dr*) învață un set de parametri αşi β, care procesează intrarea 3D după cum urmează:

$$\begin{bmatrix} \frac{\alpha_{1,i}*s_i+\beta_{1,i}*c_{1,i}*r_{1,i}}{2} & \frac{\alpha_{2,i}*s_i+\beta_{2,i}*c_{2,i}*r_{2,i}}{2} & \frac{\alpha_{3,i}*s_i+\beta_{3,i}*c_{3,i}*r_{3,i}}{2} \\ \frac{\alpha_{8,i}*s_i+\beta_{8,i}*c_{8,i}*r_{8,i}}{2} & s_i & \frac{\alpha_{4,i}*s_i+\beta_{4,i}*c_{4,i}*r_{4,i}}{2} \\ \frac{\alpha_{7,i}*s_i+\beta_{7,i}*c_{7,i}*r_{7,i}}{2} & \frac{\alpha_{6,i}*s_i+\beta_{6,i}*c_{6,i}*r_{6,i}}{2} & \frac{\alpha_{5,i}*s_i+\beta_{5,i}*c_{7,i}*r_{5,i}}{2} \end{bmatrix}$$

Numărul de parametri utilizați de stratul CSF pentru fiecare pereche de centroid este 16, generând astfel $16 \times N$ parametri de antrenat, unde N este numărul total de inductori. Straturile de pooling procesează în cele din urmă ieșirea stratului CSF, generând astfel o singură valoare pentru fiecare grup de centroizi și, astfel, produce o

matrice de dimensiune similară cu intrarea dinainte de etapa de pre-procesare. Testăm două setări diferite pentru procesarea datelor. În prima configurare, notată *8S*, toate cele 8 valori cele mai similare ale inductorului sunt populate, în timp ce în cea de-a doua configurare, notată *4S*, doar 4 cele mai similare sunt populate, restul fiind populate cu zerouri.

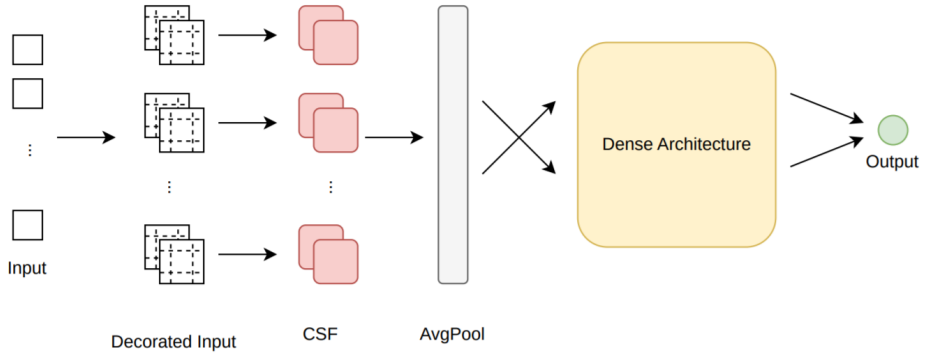


*Figura 3.5.3 Arhitectura rețelei DeepFusion Cross-Space-Fusion (DF-CSF). Aici sunt reprezentate stadiul de decorare a intrării, stratul de procesare CSF, stratul de pooling și arhitectura densă finală.*

*Protocol de antrenare*

Propunem câțiva pași esențiali în dezvoltarea acestei abordări de fuziune târzie. Primul pas constă în adunarea tuturor vectorilor individuali pentru fiecare dintre probele *M* din setul de antrenament. Căutăm apoi cea mai performantă arhitectură densă utilizând setarea prezentată în „Rețele dense" în ceea ce privește numărul de straturi, numărul de neuroni pe strat și utilizarea normalizării. Rezultatele sunt testate în funcție de setul de validare. Cea mai performantă arhitectură densă este apoi augmentată cu straturi convoluționale în al treilea pas și cu straturi Cross-Space-Fusion în al patrulea pas. Intrarea este modificată pentru utilizarea straturilor convoluționale și CSF, așa cum este descris în „Decorarea intrărilor".

### 3.5.5 Setare experimentală

Testăm metodele propuse pe mai multe tipuri de seturi de date: aceste seturi de date vizează regresia cu o clasă, regresia cu mai multe clase și sarcini de predicție multi-etichetă. Am implementat metodele pe următoarele seturi de date: MediaEval 2017 Predicting Media Interestingness [22] împărțit într-o sarcină de imagine (notat INT2017.Image) și o sarcină video (INT2017.Video), MediaEval 2015 Violent Scenes Detection [23] (VSD2015.Video), MediaEval 2018 Predicting Emotional Impact of Movies [68] împărțit în componentele de excitare (Aro2018.Video), valență (Val2018.Video) și detectarea fricii (Fear2018.Video) și, în cele din urmă, ImageCLEFmed 2019 Concept Detection [69] (Capt2019. Image).

### 3.5.6 Rezultate experimentale

| Set de date | ME top | SOA top | Emb | DF-Dens | DF-Conv |
|---|---|---|---|---|---|
| INT2017.Image (MAP@10) | 0.1385 | 0.156 | 0.1674 | 0.3355 | 0.3436 |
| INT2017.Video (MAP@10) | 0.0827 | 0.093 | 0.1129 | 0.2677 | 0.2799 |
| VSD2015.Video (MAP) | 0.296 | 0.303 | 0.392 | 0.6341 | 0.6471 |

*Tabel 2.5.1 Rezultate pentru seturile de date INT2017.Image, INT2017.Video și VSD2015.Video pentru arhitecturile dense și convoluționale.*

| Set de date | ME top | Emb | DF-Dens | DF-CSF |
|---|---|---|---|---|
| Aro2018.Video (MSE) | 0.1334 | 0.1253 | 0.0549 | 0.0543 |
| Aro2018.Video (PCC) | 0.3358 | 0.3828 | 0.8315 | 0.8422 |
| Val2018.Video (MSE) | 0.0837 | 0.0769 | 0.0626 | 0.0625 |
| Val2018.Video (PCC) | 0.3047 | 0.3972 | 0.8101 | 0.8123 |
| Fear2018.Video (IoU) | 0.1575 | 0.1733 | 0.2129 | 0.2242 |
| Capt2019.Image (F1) | 0.2823 | 0.2846 | 0.374 | 0.3912 |

*Tabel 2.5.2 Rezultate pentru seturile de date Aro2018.Video, Val2018.Video, Fear2018.Video și Capt2019.Image pentru arhitecturile dense și CSF.*

Așa cum se arată în tabelele 2.5.1 și 2.5.2, rezultatele acestor arhitecturi propuse au depășit în mod clar nu numai rezultatele de top din literatură (MEtop, SOAtop), ci și un set de metode tradiționale de fuziune (Emb). Aceste rezultate reprezintă o îmbunătățire semnificativă față de sistemele de ultimă generație, atingând chiar și până la 200,9% îmbunătățire în cazul INT2017.Video.

# Capitolul 4 - Concluzii generale și perspective

## 4.1 Contribuții și publicații

În acest capitol voi rezuma principalele contribuții personale la lucrările de cercetare publicate în timpul programului meu de cercetare doctorală. Aceste contribuții sunt după cum urmează:

- *Capitole de carte*

**C1** C.-H. Demarty, M. Sjöberg, **M.G. Constantin**,, N.Q.K. Duong, B. Ionescu, T.-T. Do, H. Wang : Predicting Interestingness of Visual Content. In book Visual Content Indexing and Retrieval with Psycho-Visual Models, Springer Multimedia Systems and Applications, Eds. J. Benois-Pineau, P. Le Callet, 2017.

**C2** B. Ionescu, H. Müller, R. Péteri. D.-T. Dang-Nguyen, ... , M. Dogariu, L.-D. Ștefan, **M.G. Constantin** : ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications. In Springer Lecture Notes in Computer Science, 12036, pp. 533-541, ISBN: 978-3-030-45441-8, DOI: https://doi.org/10.1007/978-3-030-45442-5_69, ECIR 2020 Proceedings, April 14-17, Lisbon, Portugal, 2020.

- *Reviste*

**J1** Y. Deldjoo, M.F. Dacrema, **M.G. Constantin**, H. Eghbal-zadeh, S. Cereda, M. Schedl, B. Ionescu, P. Cremonesi : Movie genome: alleviating new item cold start in movie recommendation. User Modeling and User-Adapted Interaction, ISSN 1573-1391, DOI https://doi.org/10.1007/s11257-019-09221-y, February 2019. *(Q1 journal article, Impact Factor: 4.682).*

**J2 M.G. Constantin**, M. Redi, G. Zen, B. Ionescu : Computational Understanding of Visual Interestingness Beyond Semantics: Literature Survey and Analysis of Covariates. ACM Computing Surveys, 52(2), ISSN 0360-0300, DOI http://doi.acm.org/10.1145/3301299, March 2019. *(Q1 journal article, Impact Factor: 7.990).*

**J3 M.G. Constantin**, L.D. Stefan, B. Ionescu, C.-H. Demarty, M. Sjöberg, M. Schedl, G. Gravier : Affect in Multimedia: Benchmarking Violent Scenes Detection. IEEE Transactions on Affective Computing, DOI http://dx.doi.org/10.1109/TAFFC\-.2020.2986969, April 2020. *(Q1 journal article, Impact Factor: 7.512).*

**J4** Paper under revision: **M.G. Constantin**, L.-D. Ștefan, B. Ionescu, N.Q.K. Duong, C.-H. Demarty, M. Sjöberg : Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision. *(Q1 journal article, Impact Factor: 5.698).*

- *Conferințe*

**C1** B. Boteanu, **M.G. Constantin**, B. Ionescu : LAPI @ 2016 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversification Perspective. In Working Notes Proceedings of the MediaEval 2016 Workshop, CEUR-WS.org., ISSN 1613-0073. Hilversum, The Netherlands, October 20-21, 2016.

**C2 M.G. Constantin**, B. Boteanu, B. Ionescu : LAPI at MediaEval 2016 Predicting Media Interestingness Task. In Working Notes Proceedings of the MediaEval 2016 Workshop, CEUR-WS.org., ISSN 1613-0073. Hilversum, The Netherlands, October 20-21, 2016.

**C3 M.G. Constantin**, B. Ionescu : Content Description for Predicting Image Interestingness. IEEE International Symposium on Signals, Circuits and Systems – ISSCS, July 13-14, Iași, Romania, 2017. ISI indexed conference.

**C4** B. Boteanu, **M.G. Constantin**, B. Ionescu : LAPI @ 2017 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversification Perspective. In Working Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017.

**C5 M.G. Constantin**, B. Boteanu, B. Ionescu : LAPI at MediaEval 2017 - Predicting Media Interestingness. In Working Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017.

**C6** C.A. Mitrea, **M.G. Constantin**, L.D. Stefan, M. Ghenescu, B. Ionescu : Little-Big Deep Neural Networks for Embedded Video Surveillance. IEEE International Conference on Communications – COMM, June 14-16, Bucharest, Romania, 2018. ISI indexed conference.

**C7** Y. Deldjoo, **M.G. Constantin**, M. Schedl, B. Ionescu, P. Cremonesi : MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval. ACM Multimedia Systems Conference – MMSys, June 12-15, Amsterdam, Netherlands, 2018. ISI indexed conference.

**C8** S.V. Carata, **M.G. Constantin**, V. Ghenescu, M. Chindea, M.T. Ghenescu : Innovative Multi PCNN Based Network for Green Area Monitoring - Identification and Description of Nearly Indistinguishable Areas. In Hyperspectral Satellite Images, IEEE International Geoscience and Remote Sensing Symposium - IGARSS, Valencia, Spain, 2018. ISI indexed conference.

**C9** Y. Deldjoo, **M.G. Constantin**, H. Eghbal-Zadeh, B. Ionescu, M. Schedl, P. Cremonesi : Audio-visual Encoding of Multimedia Content for Enhancing Movie Recommendations. ACM Conference Series on Recommender Systems - RecSys, October 2-7, Vancouver, Canada, 2018. ISI indexed conference.

**C10** Y. Deldjoo, **M.G. Constantin**, A. Dritsas, B. Ionescu, M. Schedl : The MediaEval 2018 Movie Recommendation Task: Recommending Movies Using Content. In Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, October 29-31, 2018.

**C11 M.G. Constantin**, B. Ionescu, C.-H. Demarty, N.Q.K. Duong, X. Alameda-Pineda, M. Sjöberg : The Predicting Media Memorability Task at MediaEval 2019. In Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, October 27-29, 2019.

**C12 M.G. Constantin**, C. Kang, G. Dinu, F. Dufaux, G. Valenzise, B. Ionescu : Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, October 27-29, 2019.

**C13** M. Dogariu, L.-D. Ștefan, **M.G. Constantin**, B. Ionescu : Human-Object Interaction: Application to Abandoned Luggage Detection in Video Surveillance Scenarios. IEEE International Conference on Communications - COMM, June 18-20, Bucharest, Romania, 2020. ISI indexed conference.

**C14** L.-D. Ștefan, Ș. Abdulamit, M. Dogariu, **M.G. Constantin**, B. Ionescu : Deep Learning-based Person Search with Visual Attention Embedding. IEEE International Conference on Communications - COMM, June 18-20, Bucharest, Romania, 2020. ISI indexed conference.

**C15** L.-D. Ștefan, **M.G. Constantin**, B. Ionescu : System Fusion with Deep Ensembles. ACM International Conference on Multimedia Retrieval - ICMR, October 26-29, Dublin, Ireland, 2020. ISI indexed conference.

**C16 M.G. Constantin**, L.-D. Ștefan, B. Ionescu: DeepFusion: Deep Ensembles for Domain Independent System Fusion. International Conference on Multimedia Modeling - MMM, June 22-24, Prague, Czech Republic, 2021. ISI indexed conference.

În (C2) am propus implementarea unui set de descriptori vizuali tradiționali pentru prezicerea interesului în multimedia. Validarea experimentală se efectuează pe setul de date MediaEval 2016 Predicting Media Interestingness.

În (C3) și (C5) am propus implementarea unui set mare de trăsături estetice, bazate pe reguli de culoare, textură, fotografie și compoziție, pentru a prezice interesul media. Metodele sunt validate atât în versiunea 2016, cât și în versiunea 2017 ale seturilor de date MediaEval Predicting Media Interestingness, precum și implementarea schemelor de fuziune timpurie și târzie pentru optimizarea performanței. Rezultatele înregistrate pe setul de imagine 2016 reprezintă încă cel mai performant sistem în ceea ce privește performanța MAP.

În (J1), (C7), (C9), (C10) am propus implementarea descriptorilor vizuali pentru crearea sistemelor de recomandare a filmelor. Aceste lucrări de cercetare au produs și setul de date MMTF-14K, unde am furnizat un set de descriptori estetici și bazați pe DNN ca sisteme de bază pentru cercetătorii care doresc să utilizeze setul nostru de date.

(J2) reprezintă în prezent, cel mai mare studiu al literaturii privind predicția interesului multimedia și a covariabilelor sale. Contribuțiile mele la această lucrare sunt legate de studiul abordărilor de viziune pe computer pentru predicția gradului de interes și conceptele sale corelate, crearea unui model de taxonomie care studiază corelațiile pozitive, negative și încă neexplorate dintre interes și alte concepte subiective și, cu un grad mai mic de implicare, studiul înțelegerii umane asupra interesului.

În (C11) am fost principalul organizator al competiției MediaEval 2019 Predicting Media Memorability, cu contribuții în ajutorarea participanților MediaEval, evaluarea sistemelor trimise și teoretizarea tendințelor generale în ceea ce privește cele mai bune practici.

În (C12) am propus implementarea DNN-urilor bazate pe recunoașterea acțiunii pentru predicția memorabilității media. Rezultatele sunt validate pe MediaEval 2019 Predicting Media Interestingness, iar schemele de fuziune timpurie și târzie sunt implementate pentru optimizarea performanței.

(J3) reprezintă o analiză aprofundată a setului de date VSD96, care vizează detectarea scenelor video violente. Principalele mele contribuții la această lucrare sunt reprezentate de analiza generală a metodelor utilizate pe acest set de date de către un număr mare de autori, un studiu al influenței descriptorilor asupra rezultatelor predicției și formularea unor concluzii principale cu privire la predicția violenței.

(J4), o lucrare în curs de examinare, reprezintă o analiză aprofundată a setului de date Interestingness10k, care vizează predicția gradului de interes al imaginilor și al videoclipurilor. Principalele mele contribuții la această lucrare sunt următoarele: analiza performanței generale a sistemelor care utilizează acest set de date, o analiză a influenței descriptorilor asupra performanței sistemelor, studiul capacităților de

generalizare a sistemelor și recomandări cu privire la creșterea performanței sistemelor. Unele contribuții comune includ: studiul abordărilor DNN de ultimă generație și interpretabilitatea rezultatelor, precum și implementarea sistemelor statistice, de boosting și de fuziune târzie bazate pe rețele adânci pentru îmbunătățirea rezultatelor înregistrate în timpul MediaEval 2016 și 2017 ale Predicting Media Interestingness.

(C15) reprezintă o abordare nouă în ceea ce privește sistemele de fuziune târzie. Noutatea este reprezentată de introducerea arhitecturilor DNN ca principală metodă de asamblare pentru combinarea ieșirilor de predicție ale inductorilor. Principalele mele contribuții la această lucrare sunt reprezentate de crearea unei metode de decorare a intrărilor, care facilitează o grupare spațială a ieșirilor similare și de implementarea straturilor convoluționale pentru procesarea intrării decorate. Validarea se efectuează pe trei seturi de date de regresie, și anume procesarea de imagine și video din MediaEval 2017 Predicting Media Interestingness și detectarea scenelor violente din MediaEval 2015 și, după cum arată rezultatele, aceste metode îmbunătățesc foarte mult performanțele sistemelor.

(C16) prezintă un alt set de abordări noi în ceea ce privește sistemele de fuziune. Păstrând abordarea arhitecturală bazată pe DNN, noutatea acestei lucrări este reprezentată de introducerea unui strat DNN special conceput pentru acest tip de sarcină, stratul Cross-Space-Fusion. Principalele mele contribuții la această lucrare sunt reprezentate de crearea unei alte metode de decorare a intrărilor și de crearea și dezvoltarea stratului CSF. Validarea se efectuează pe o varietate de sarcini care acoperă diferite condiții de validare: regresie în două clase (reprezentată de detecția Arousal și Valence din MediaEval 2018 Emotional Impact of Movies), clasificare binară (reprezentată de detectare a fricii din setul MediaEval 2018 Emotional Impact of Movies) și clasificarea multi-etichetă (reprezentată de setul ImageCLEF 2019 Medical Concept Detection).

# 4.2 Concluzii

Această teză prezintă contribuțiile personale la analiza automată a impactului vizual al datelor multimedia, cu accent pe studiul gradului de interes, esteticii, memorabilității, violenței și valorii afective și emoțiilor. Capitolul 2 prezintă o analiză a stadiului actual din literatura în ceea ce privește taxonomia conceptelor și definițiile, teoriile privind înțelegerea umană a proprietăților multimedia subiective, seturi de date și studii ale utilizatorilor, abordări computaționale și aplicații actuale și perspective viitoare privind utilizarea dintre aceste proprietăți. Capitolul 3 prezintă contribuțiile mele la acest domeniu. Prima parte a acestui capitol acoperă seturile de date și inițiativele de benchmarking la care am contribuit. În continuare, teza prezintă mai multe metode de predicție automată dezvoltate în timpul programului meu de doctorat și analizează contribuțiile în acest domeniu adus de aceste metode.

Metodele prezentate aici sunt legate de: (i) predicția gradului de interes media prin descriptori vizuali tradiționali printr-o metoda SVM și implementarea descriptorilor bazați pe estetică și a schemelor statistice de fuziune târzie pentru

predicția gradului de interes; (ii) detectarea scenelor violente prin implementarea unei abordări ConvLSTM; (iii) predicția memorabilității media cu ajutorul rețelelor neuronale adânci de recunoaștere a acțiunii; (iv) crearea unei noi abordări bazate pe învățarea adâncă de fuziune târzie, crearea de noi metode de decorare a intrărilor care să permită prelucrarea inductorilor corelați în sistemele de fuziune profundă și un nou tip de strat de rețea neuronală, Cross-Space-Fusion, special conceput pentru procesarea sistemelor în fuziunea târzie.

## 4.3 Perspective de viitor

În continuarea acestei lucrări, cel mai important aspect ar fi implementarea sistemelor care sunt mai bine adaptate pentru sarcinile lor respective. Unele exemple sunt deja prezentate în această teză, adică un set de caracteristici estetice, dar consider că, prin implementarea mai multor tipuri de sisteme bazate pe lucrări anterioare din domeniile psihologiei și analizei comportamentului, pot fi construite arhitecturi mai bune și rezultatele acestora ar fi benefice comunității multimedia.

Mai mult, având în vedere rezultatele sistemului de fuziune târzie, consider că acesta reprezintă o direcție de cercetare foarte interesantă pentru viitor. În timp ce această abordare reprezintă prima încercare de a crea astfel de sisteme de fuziune bazate pe rețele neuronale, dezvoltările viitoare pot include: crearea unor metode noi de decorare a intrărilor, adăugarea de noi straturi și scheme de instruire pentru straturile existente și studii în ceea ce privește optimizarea colecției de inductori utilizați.

# Bibliografie

[1] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huber-man. Influence and passivity in social media. In Machine Learning and Knowledge Discovery in Databases, volume 6913, 18–33. Springer Berlin Heidelberg, 2011.
[2] Berlyne, D. E. (1949). Interest as a psychological concept. British Journal of Psychology, 39(4), 184.
[3] Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. Emotion, 5(1), 89.
[4] Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. The role of interest in learning and development, 11, 213-214.
[5] Zangwill, N. (2003). Aesthetic judgment.
[6] Constantin, M. G., Ionescu, B., Demarty, C. H., Duong, N. Q., Alameda-Pineda, X., & Sjöberg, M. (2019, October). Predicting Media Memorability Task at MediaEval 2019. In Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France.
[7] Demarty, C. H., Penet, C., Schedl, M., Bogdan, I., Quang, V. L., & Jiang, Y. G. (2013, October). The mediaeval 2013 affect task: violent scenes detection. In MediaEval 2013 Working Notes (p. 2).
[8] Cabanac, M. (2002). What is emotion?. Behavioural processes, 60(2), 69-83.
[9] Mehrabian, A. (1980). Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies (Vol. 2). Cambridge, MA: Oelgeschlager, Gunn & Hain.
[10] Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.
[11] Berlyne, D. E. (1960). Conflict, arousal, and curiosity.
[12] Silvia, P. J. (2009). Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. Psychology of Aesthetics, Creativity, and the Arts, 3(1), 48.
[13] Izard, C. E. (1984). Emotion-cognition relationships and human. Emotions, cognition, and behavior, 17.
[14] Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?. Personality and social psychology review, 8(4), 364-382.
[15] Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006, May). Studying aesthetics in photographic images using a computational approach. In European conference on computer vision (pp. 288-301). Springer, Berlin, Heidelberg.
[16] Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. Journal of verbal Learning and verbal Behavior, 6(1), 156-163.
[17] Arendt, H. (1970). On violence. Houghton Mifflin Harcourt.
[18] Galtung, J. (1990). Cultural violence. Journal of peace research, 27(3), 291-305.
[19] Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. Journal of experimental psychology: General, 123(4), 394.
[20] Chen, C. M., & Sun, Y. C. (2012). Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. Computers & Education, 59(4), 1273-1285.

[21] Soleymani, M. (2015, October). The quest for visual interest. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 919-922).
[22] Demarty, C. H., Sjöberg, M., Ionescu, B., Do, T. T., Gygli, M., & Duong, N. (2017, September). Mediaeval 2017 predicting media interestingness task. In MediaEval workshop.
[23] Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., ... & Chen, L. (2015, September). The MediaEval 2015 Affective Impact of Movies Task. In MediaEval.
[24] Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1633-1640).
[25] Jiang, Y. G., Wang, Y., Feng, R., Xue, X., Zheng, Y., & Yang, H. (2013, June). Understanding and predicting interestingness of videos. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 27, No. 1).
[26] Demarty, C. H., Sjoberg, M., Ionescu, B., Do, T. T., Wang, H., Duong, N. Q., & Lefebvre, F. (2016). Mediaeval 2016 predicting media interestingness task. In MediaEval 2016 Workshop.
[27] Ke, Y., Tang, X., & Jing, F. (2006, June). The design of high-level features for photo quality assessment. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 1, pp. 419-426). IEEE.
[28] Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. IEEE Journal of selected topics in Signal Processing, 3(2), 236-252.
[29] Parikh, D., Isola, P., Torralba, A., & Oliva, A. (2012). Understanding the intrinsic memorability of images. Journal of Vision, 12(9), 1082-1082.
[30] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[31] Fajtl, J., Argyriou, V., Monekosso, D., & Remagnino, P. (2018). Amnet: Memorability estimation with attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6363-6372).
[32] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., & Theodoridis, S. (2010, May). Audio-visual fusion for detecting violent scenes in videos. In Hellenic conference on artificial intelligence (pp. 91-100). Springer, Berlin, Heidelberg.
[33] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012, June). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-6). IEEE.
[34] Hanson, A., Pnvr, K., Krishnagopal, S., & Davis, L. (2018). Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.
[35] Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T. S., & Sun, X. (2014, November). Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 47-56).
[36] Jou, B., Chen, T., Pappas, N., Redi, M., Topkara, M., & Chang, S. F. (2015, October). Visual affect around the world: A large-scale multilingual visual sentiment ontology. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 159-168).
[37] Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. Journal of experimental psychology: General, 123(4), 394.
[38] Constantin, M. G., Ştefan, L. D., Ionescu, B., Duong, N. Q., Demarty, C. H., & Sjöberg, M. (2021). Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision, 1-25.

[39] Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C. H., Sjoberg, M., Schedl, M., & Gravier, G. (2020). Affect in multimedia: Benchmarking violent scenes detection. IEEE Transactions on Affective Computing.
[40] Deldjoo, Y., Constantin, M. G., Ionescu, B., Schedl, M., & Cremonesi, P. (2018, June). MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 450-455).
[41] Sudhakaran, S., Escalera, S., & Lanz, O. (2020). Gate-shift networks for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1102-1111).
[42] Shen, Y, Demarty, C. H., Duong, N. Q. K. Technicolor@MediaEval 2016 Predicting Media Interestingness Task
[43] Vasudevan, A. B., Gygli, M., Volokitin, A., & Van Gool, L. (2016, October). ETH-CVL@ MediaEval 2016: Textual-Visual Embeddings and Video2GIF for Video Interestingness. In MediaEval.
[44] Constantin, M. G., Boteanu, B., & Ionescu, B. (2016, October). LAPI at MediaEval 2016 Predicting Media Interestingness Task. In MediaEval.
[45] Constantin, M. G., Boteanu, B. A., & Ionescu, B. (2017). LAPI at MediaEval 2017-Predicting Media Interestingness. In MediaEval.
[46] Constantin, M. G., & Ionescu, B. (2017, July). Content description for Predicting image Interestingness. In 2017 International Symposium on Signals, Circuits and Systems (ISSCS) (pp. 1-4). IEEE.
[47] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
[48] Van De Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009). Learning color names for real-world applications. IEEE Transactions on Image Processing, 18(7), 1512-1523.
[49] Haas, A. F., Guibert, M., Foerschner, A., Calhoun, S., George, E., Hatay, M., ... & Rohwer, F. (2015). Can we measure beauty? Computational evaluation of coral reef aesthetics. PeerJ, 3, e1390.
[50] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. arXiv preprint arXiv:1506.04214.
[51] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[52] Sudhakaran, S., & Lanz, O. (2017, August). Learning to detect violent videos using convolutional long short-term memory. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE.
[53] Constantin, M. G., Kang, C., Dinu, G., Dufaux, F., Valenzise, G., & Ionescu, B. (2019, October). Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In MediaEval 2019 Workshop.
[54] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[55] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36). Springer, Cham.

[56] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).
[57] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).
[58] Kang, C., Valenzise, G., & Dufaux, F. (2019, September). Predicting Subjectivity in Image Aesthetics Assessment. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.
[59] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
[60] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
[61] Ştefan, L. D., Constantin, M. G., & Ionescu, B. (2020, June). System Fusion with Deep Ensembles. In Proceedings of the 2020 International Conference on Multimedia Retrieval (pp. 256-260).
[62] Constantin, M. G., Ştefan, L. D., & Ionescu, B. (2021, June). DeepFusion: Deep Ensembles for Domain Independent System Fusion. In International Conference on Multimedia Modeling (pp. 240-252). Springer, Cham.
[63] Wang, S., Chen, S., Zhao, J., & Jin, Q. (2018, October). Video interestingness prediction based on ranking model. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data (pp. 55-61).
[64] Azcona, D., Moreu, E., Hu, F., Ward, T. E., & Smeaton, A. F. (2020, September). Predicting media memorability using ensemble models. CEUR Workshop Proceedings.
[65] Sun, J. J., Liu, T., & Prasad, G. (2019). Gla in mediaeval 2018 emotional impact of movies task. arXiv preprint arXiv:1911.12361.
[66] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.
[67] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
[68] Dellandréa, E., Huigsloot, M., Chen, L., Baveye, Y., Xiao, Z., & Sjöberg, M. (2018). The MediaEval 2018 Emotional Impact of Movies Task. In Multimedia Benchmark Workshop. CEUR.
[69] Pelka, O., Friedrich, C. M., Seco De Herrera, A. G., & Müller, H. (2019, July). Overview of the ImageCLEFmed 2019 concept detection task. CEUR Workshop Proceedings.