

# Automatic analysis of the visual impact of multimedia data

*by* Mihai Gabriel Constantin

---

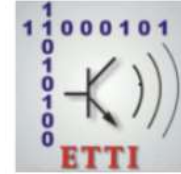
**Submission date:** 23-Oct-2020 03:19PM (UTC+0300)

**Submission ID:** 1424159736

**File name:** TEZA\_\_\_Mihai\_Gabriel\_Constantin\_\_\_FINAL\_2.pdf (1.53M)

**Word count:** 28710

**Character count:** 161049



“POLITEHNICA” UNIVERSITY OF BUCHAREST

ETTI-B DOCTORAL SCHOOL

Decision No. 569 from 25.09.2020

# Automatic analysis of the visual impact of multimedia data

Analiza automată a impactului vizual al datelor  
multimedia

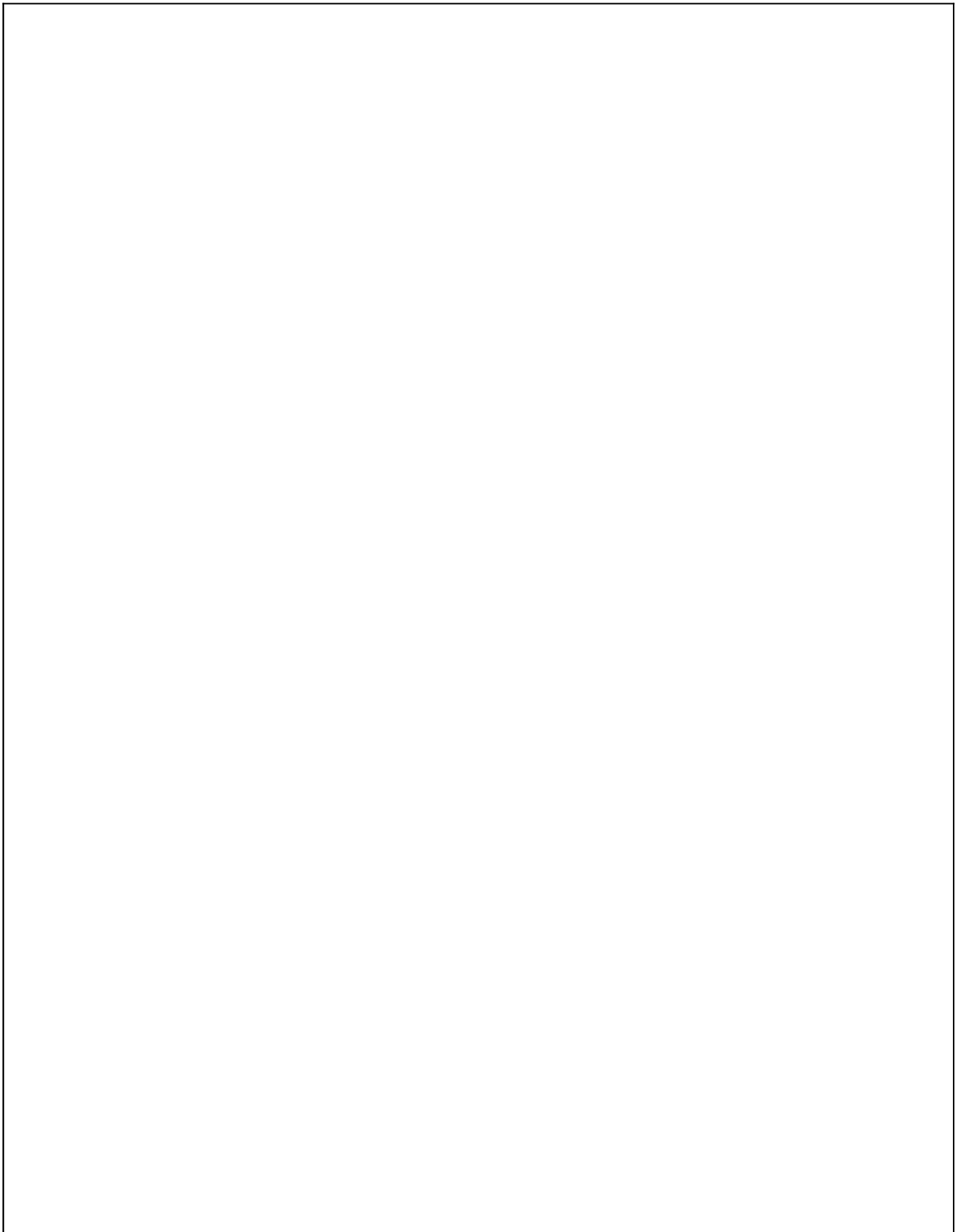
by Mihai Gabriel Constantin

<sup>31</sup> A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
in Electronics and Telecommunications

## <sup>13</sup> COMISIA DE DOCTORAT

Președinte	Prof. Dr. Ing. Gheroghe Brezeanu	de la	Universitatea Politehnica București
Conducător de doctorat	Prof. Dr. Ing. Bogdan Ionescu	de la	Universitatea Politehnica București
Referent	Prof. Dr. Ing. Martha Larson	de la	Radboud University Olanda
Referent	Dr. Ing. Claire-Hélène Demarty	de la	InterDigital, Franța
Referent	Prof. Dr. Ing. Mihai Ciuc	de la	Universitatea Politehnica București

Bucharest 2020



## Acknowledgements

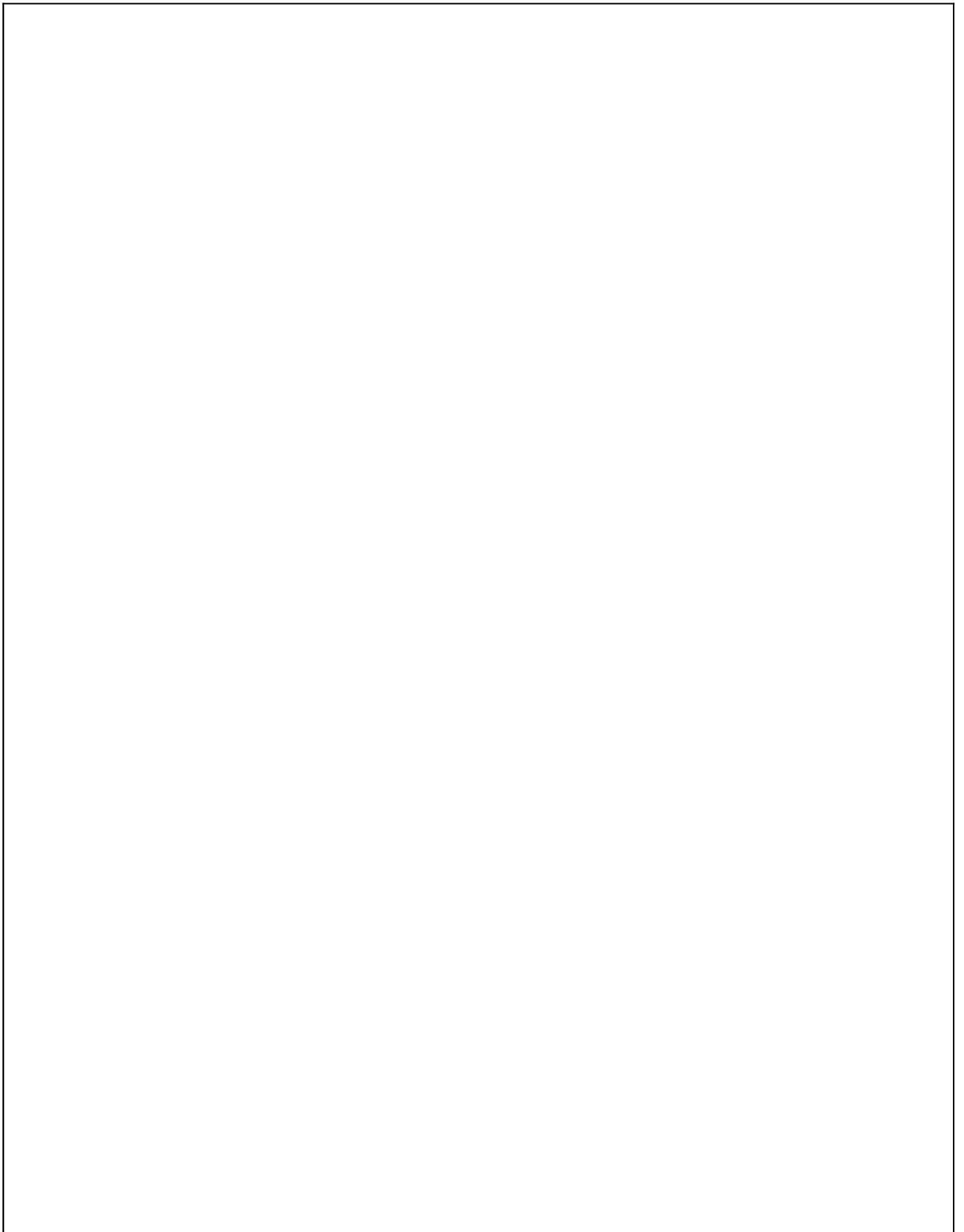
I would firstly like to thank my thesis coordinator, Prof. Dr. Ing. Bogdan Ionescu, for his patience and guidance<sup>41</sup> during the time I spent working on this thesis. I would also like to thank him for introducing me to this domain and to the scientific community that revolves around it. With his help, I discovered the challenging, ever-evolving, and fascinating world of academia and research. As my doctoral program reaches this stage, I can only hope for a long and fruitful future collaboration with him.

I would also like to thank the MediaEval community and the person who drives and organizes this community, Prof. Dr. Ing. Martha Larson. The benchmarking tasks published in this community represent one of the pillars upon which this thesis is created.

Also, a big thank you to my colleagues in the Multimedia Lab for their help, input, and collaboration on the research projects and paper we published together. My gratitude also goes to my friends, who encouraged me to follow a career in academia.

Finally,<sup>38</sup> I would like to thank my parents for their continuous help, support, and encouragement in all my endeavors.



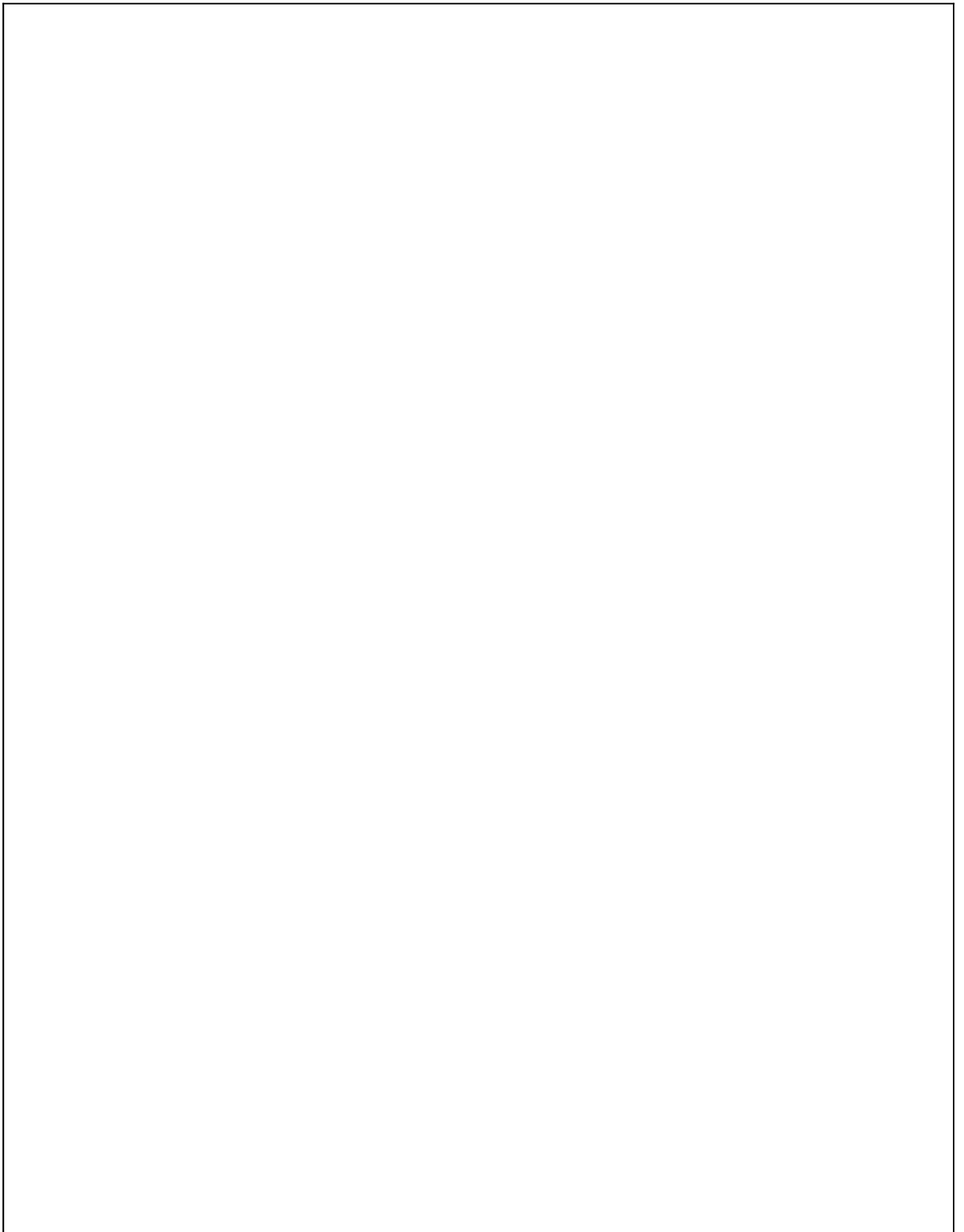


# Contents

Acknowledgements	iii
<sup>30</sup> List of abbreviations	ii
<b>1 Introduction</b>	<b>1</b>
1.1 Domain of the thesis	1
1.2 Motivation of the thesis	2
1.3 Content of the thesis	4
<b>2 Theoretical aspects</b>	<b>5</b>
2.1 Taxonomy and definitions	6
2.2 Human understanding of the subjective properties of multimedia data	10
2.3 Datasets and user studies	13
2.4 Computational approaches	17
2.4.1 Interestingness	17
2.4.2 Aesthetic value	19
2.4.3 Memorability	20
2.4.4 Violence	21
2.4.5 Affective value and emotions.	22
2.5 Applications	23
2.6 Conclusions	25

<b>3 Personal contributions</b>	<b>27</b>
3.1 Datasets and evaluation	27
3.1.1 Interestingness prediction	27
3.1.2 Violence prediction	35
3.1.3 Memorability prediction	39
3.1.4 Content recommendation	41
3.2 Predicting media interestingness	43
3.2.1 Introduction	43
3.2.2 SVM-based learning systems	43
3.2.3 Aesthetic features and late fusion learning systems	48
3.2.4 Conclusions	54
3.3 Predicting violent scenes	57
3.3.1 Introduction	57
3.3.2 Temporal deep learning systems	57
3.3.3 Conclusions	60
3.4 Predicting media memorability	61
3.4.1 Introduction	61
3.4.2 Action-based deep learning systems	61
3.4.3 Conclusions	65
3.5 Late fusion with deep ensemble systems	67
3.5.1 Introduction	67
3.5.2 Motivation	67
3.5.3 Previous work	68
3.5.4 Proposed approach	68
3.5.5 Experimental setup	75
3.5.6 Experimental results	77
3.5.7 Conclusions	81

<b>4 General conclusions and perspectives</b>	<b>83</b>
4.1 Contributions and publications . . . . .	83
4.2 Conclusions . . . . .	90
4.3 Future perspectives . . . . .	91
<b>Bibliography</b>	<b>93</b>



## List of abbreviations

1D, 2D, 3D - one-, two-, three-dimensional

BN - Batch normalization

CNN - Convolutional neural network

CSF - Cross-Space-Fusion layer

DNN - Deep neural network

HOG - Histogram of oriented gradients

HSV - Hue-saturation-value

IQR - Interquartile range

KF - K-fold

kNN - k-nearest neighbors algorithm

LBP - Local binary patterns

LOF - Local outlier factor

LSTM - Long short-term memory

MAP - Mean average precision

<sup>66</sup>MLP - Multi-layered perceptron

RBF - Radial basis function

<sup>29</sup>SIFT - Scale-invariant feature transform

SVM - Support vector machine

VAD - Valence-arousal-dominance

# Chapter 1

## Introduction

### 1.1 Domain of the thesis

This thesis presents and analyzes several aspects and <sup>65</sup>state of the art methods that cover the automatic analysis of the visual impact of multimedia data, with an accent on the study of a number essential concepts in this domain, such as interestingness, aesthetics, memorability, violence, and affective value and emotions. While more traditional computer vision tasks attempt to solve problems that have objective ground-truth values that all or most annotators would agree with, such as object detection or scene classification, recent developments in deep neural network processing, social media, hardware availability and cost, psychological studies, and big data availability allowed scientists to expand their research into domains that target more subjective concepts. In the latter case, ground truth may depend on a large number of human-centric factors, including, but not limited to, personal preferences, cultural background, cognitive abilities, and current psychological state. Predicting and understanding such concepts with the help of computer vision methods dramatically increases the utility and added value created by implementing such methods, allowing scientists and developers to predict how multimedia data affects viewers.

However, the development of such methods is not trivial. Researchers from many different domains must be involved and must work together in order to create accurate predictors that can function in a mostly online, real-world environment that deals with large amounts of diverse visual data. Researchers in cognitive and humanities sciences, physiologists, specialists in human behavior, human data annotation, and computer vision algorithm developers must come together in order to define these concepts, create theories about how they influence perception and behavior, collect and annotate a large amount of data and create the computer vision methods that can predict the concepts.

In this thesis, I present a literature survey on my main concepts of interest for my field of study, which predominantly revolves around *interestingness*, *aesthetic value*, *memorability*, *violence* and *affective value and emotional content*, and continue with presenting my main contributions, both to the collection of datasets and the creation of common evaluation benchmarks, to computer vision methods for the prediction of such concepts, as well as the creation of a novel deep learning-based late fusion system, that significantly increases the performance of its inducer systems. All these contributions are developed during my Ph.D. studies.

## 1.2 Motivation of the thesis

This thesis aims to contribute to the understanding of such subjective concepts, study, discover, and underline the current best practices and best-performing methods and models for certain tasks, and create computer vision methods that successfully predict the targeted concepts. Given the advent of social media, increasingly larger collections of images and videos are available for users, and it becomes increasingly difficult to navigate them. On the positive side, access to a larger amount of data can be beneficial for system development, as more training and testing samples are



available, especially for deep neural networks, that are known for their high demand for annotated data. While, as previously shown, this extensive collection of concepts present varying degrees of subjectivity, and, therefore, inter- and even intra-rater reliability with regards to the annotated image and video samples in given datasets can significantly vary, the interest for computer vision methods that solve these problems and predict these concepts is growing, regardless of the difficulties created by the inherent concept subjectivity. From this perspective, there is a large demand for these methods, mostly driven by social media, media sharing, advertising, and media archiving platforms. These particular branches of industry would benefit from the creation of automatic predictors, recommender systems based on these concepts, automatic filters, and other functionalities that would be impossible to implement without the help of computer vision, machine learning, and artificial intelligence.

Currently, some of these concepts are starting to be implemented in professional solutions and web services. Pioneers in this direction are represented by popular websites and social media platforms, like Flickr<sup>[1]</sup>, who implement a social interestingness-based metric for creating suggestions with regards to new images and posts, or Google Photos<sup>[2]</sup> that can create short summaries of photos based on the appeal of photos uploaded by users in their personal collections. Support from the industry is also manifested by the support for particular tasks that aid both the research community and the industry, such as InterDigital's support for the study of multimedia interestingness, memorability, violence, and emotional content prediction<sup>[3]</sup>. Thus, researchers that create tasks, datasets, and computational models are motivated to keep in mind and target realistic use case scenarios that can be implemented in such environments.

<sup>1</sup><https://www.flickr.com/>

<sup>2</sup><https://www.google.com/photos/about/>

<sup>3</sup><https://www.interdigital.com/datasets/>

### 1.3 Content of the thesis

The rest of this thesis is divided into 3 Chapters. The first one presents the current state-of-the-art with regards to taxonomies, psychological studies, datasets, user studies, and computational approaches developed by researchers from different domains that handle the problem of defining and predicting the subjective proprieties of multimedia data. The second chapter presents personal contributions to this domain, with regards to the datasets and evaluation benchmarks I helped create, and to original computational methods and models for the prediction of some of these concepts, as well as a generalized deep learning-based collection of late fusion approaches that can accurately predict the given concepts, using a large selection of weaker input inducers. The thesis ends with some general conclusions and perspectives for future works, as well as a summary of my papers and contribution to those papers.

## Chapter 2

### Theoretical aspects

In today's internet and big data landscape, users are constantly bombarded with large quantities of multimedia data, sometimes creating that data themselves via personal photo collections, social media posts, or personal vlogs. It is indeed difficult to keep track of all that information. Researchers have shown that this constant feed of information, both visual and otherwise, can significantly reduce the human attention span [138]. This environment creates the need for the development of systems that would help human users navigate this tremendous amount of data, whether we are talking about systems aimed at sorting data, based on how interesting, appealing, or memorable it is, systems aimed at creating filters capable of detecting violent or emotionally scarring data, or recommending other media samples that are more in tuned with personal user preferences. One of the hardest challenges these systems face is represented by the definitions of these concepts, considering that, unlike more tangible tasks such as detecting an object in an image, most of the times, it is hard for human subjects to agree on what is interesting, aesthetically pleasing, violent, and so on. The subjective nature of these proprieties does make their prediction and classification one of the more challenging tasks in computer vision today. A close collaboration between theorists in humanities, human behavior, and computer

vision, is therefore necessary in order to create algorithms and market-ready systems. This chapter will present a literature review and analysis focused on concepts that will be used throughout the thesis, namely *interestingness*, *aesthetics*, *memorability*, *violence*, and *affective value and emotions*.

## 2.1 Taxonomy and definitions

As previously mentioned, the first important step in analyzing these targeted concepts is creating a list of possible definitions for them, having as starting point psychological theories, applied human studies, and use-case scenarios/ An extensive set of subjective proprieties has been studied in the current literature. As we present in [33], some taxonomies can be built in order to understand and classify these concepts. Table 2.1 presents a list of subjective concepts studied by scientists, grouped according to a central common theme. For example, novelty, originality, unexpectedness, etc., tend to measure the novelty of media samples from different perspectives and therefore fall into the same central theme.

Another possible approach regarding the creation of taxonomies is an analysis of concept correlation. In this case, having just one target concept as a starting point, a list of correlation with other concepts can be created based on research works in the current state-of-the-art literature. These correlations can be positive, negative, or undefined (or mostly not explored). Such an example is presented in our work [33] and in Table 2.2 where a taxonomy based on correlation with interestingness is presented. The table represents a thorough analysis that considers papers from psychological, user study-based, or computational perspectives. Furthermore, an example of a scientific paper that studies the correlation is given for each concept. Even in such a thorough analysis, some controversies arise that show certain concepts <sup>1</sup> to be both positively and negatively correlated with interestingness. Such issues may occur due



Table 2.1: Taxonomy. List of concepts that are covered in the current state-of-the-art literature as presented in Constantin et al. [33]. Concepts in this table are grouped according to a central theme.

	<i>Theme</i>	<i>Close concepts</i>
1	Interestingness	Interestingness
2	Affective Value and Emotions	Dimensional Emotion Space (Valence/Pleasantness, Arousal, Dominance) and Categorical Emotion Space (Happiness, Boredom, etc.)
3	Aesthetic Value	Aesthetic Value and Cuteness
4	Memorability	Memorability
5	Novelty	Novelty, Originality, Unusualness, Unexpectedness, Distinctiveness and Familiarity
6	Complexity	Complexity and Simplicity
7	Coping Potential	Coping Potential, Comprehensibility, Challenge and Uncertainty
8	Visual Composition and Stylistic Attributes	Symmetry, Balance/Harmony, Photographic Composition, Naturalness and Realism
9	Social Interestingness	Popularity and Virality
10	Creativity	Creativity
11	Humor	Humor, Irony and Sarcasm
12	Urban Perception	Urban Interestingness
13	Saliency	Saliency and Attention

to different factors, including different experimental setups in computer vision tasks, different demographic spread in user studies, differences in understanding the analyzed concept, its definition, and scope, or merely different preferences for the chosen annotators.

*Interestingness.* Berlyne [9] theorizes interest as a primary factor for human motivation and behavior and points out several defining factors of interest [11], such as novelty, in the context of information theory, pointing out that interest arises when new information is compared already existing information by human subjects. More to the point, Chamaret et al. [20] define visual interest as the power of a visual sample to induce interest in a viewer. Furthermore, Silvia et al. [153] relate interest to learning and the will to explore. Similarly, Hidi and Anderson [87] propose that personal

Table 2.2: Taxonomy. List of concepts, grouped by positive, negative or unexplored correlation with interestingness as presented in Constantin et al. [33]. Correlations are studied from a physiological or cognitive ( $t$ ), user studies-based ( $u$ ) or computational ( $c$ ) perspective. Controversies are marked with \*.

Positively correlated	Negatively correlated	Unexplored
<ul style="list-style-type: none"> <li>• Valence<sup>(u,t)*</sup> [74]</li> <li>• Arousal<sup>(u,c)</sup> [160]</li> <li>• Aesthetic Value<sup>(u,t,c)*</sup> [90]</li> <li>• Novelty<sup>(u,t,c)</sup> [74]</li> <li>• Unusualness<sup>(c)</sup> [187]</li> <li>• Unexpectedness<sup>(t)</sup> [124]</li> <li>• Complexity<sup>(u,t,c)</sup> [153]</li> <li>• Coping potential<sup>(u,t)*</sup> [155]</li> <li>• Uncertainty<sup>(t)</sup> [10]</li> <li>• Balance/Harmony<sup>(u,c)</sup> [96]</li> <li>• Naturalness<sup>(u)</sup> [79]</li> <li>• Photo Composition<sup>(c)</sup> [96]</li> <li>• Humor<sup>(t,c)</sup> [96]</li> <li>• Urban interestingness<sup>(u)</sup> [141]</li> <li>• Saliency<sup>(u)</sup> [54]</li> <li>• Attention<sup>(t)</sup> [12]</li> <li>• Popularity<sup>(u,c)*</sup> [77]</li> </ul>	<ul style="list-style-type: none"> <li>• Valence<sup>(u,t)*</sup> [173]</li> <li>• Boredom<sup>(u,t)</sup> [61]</li> <li>• Aesthetic Value<sup>(t)*</sup> [146]</li> <li>• Memorability<sup>(u)</sup> [93]</li> <li>• Coping potential<sup>(u)*</sup> [160]</li> <li>• Challenge<sup>(u)*</sup> [21]</li> <li>• Virality<sup>(u,c)</sup> [50]</li> <li>• Popularity<sup>(u,t)*</sup> [90]</li> <li>• Familiarity<sup>(u)</sup> [23]</li> </ul>	<ul style="list-style-type: none"> <li>• Dominance</li> <li>• Cuteness</li> <li>• Originality</li> <li>• Distinctiveness</li> <li>• Comprehensibility</li> <li>• Symmetry</li> <li>• Realism</li> <li>• Irony, Sarcasm</li> <li>• Creativity</li> <li>• Urban Perception</li> </ul>

preferences may be less critical in inducing interest in a person than the appeal of the activity or learning task being performed.

*Aesthetic value.* Aesthetics is mainly defined as a branch of philosophy [186], that studies the appeal and beauty of natural scenes and artistic compositions. In several user studies, authors often tend to use “pleasantness” as a descriptor or synonym of aesthetics [74, 160].

*Memorability* is defined as an intrinsic propriety of visual samples [92], that measures how likely subjects are to remember the images and videos that are presented to them. Some authors use short-term and long-term memorability [48, 47] separa-

tion in describing this visual propriety, thus recognizing that, while a video can be memorable for a short period (several minutes or hours), it can be forgotten in the long run (after several days).

*Violence.* While the concept of violence may seem less subjective than others, studies have shown that human annotators do not necessarily agree on whether a visual sample is violent or not. Several studies have used more than one definition of violence, including during the MediaEval<sup>1</sup> Violent Scenes Detection task [46], where authors proposed an “objective” definition (“physical violence or accident resulting in human injury or pain”) and a “subjective” definition (where violence is defined as images “which one would not let an eight years old child see, because they contain physical violence”).

*Affective value and emotions.* The affective value of media items is defined as their ability to induce a set of emotional responses in viewers [18]. From one perspective, they can be described in a mathematical 2D or 3D space, according to the valence-arousal-dominance axes (or only valence and arousal). The VAD space attempts to map all human emotions on these three axes, corresponding to pleasure-displeasure measuring the valence or pleasantness of the emotion, arousal-nonarousal measuring the intensity of the emotion, and dominance-submissiveness measuring the controlling nature of the emotion [118]. From another perspective, emotions can be described in a categorical space, where a set of basic emotions are identified and defined. Ekman [53] identifies a set of 6 basic emotions: “anger, disgust, fear, joy, sadness and surprise”, while Plutchik [132] considers 8 bipolar emotions: “anger-fear, joy-sadness, anticipation-surprise and trust-disgust”.

---

<sup>1</sup><http://www.multimediaeval.org/>

## 2.2 Human understanding of the subjective properties of multimedia data

The human understanding of these concepts is extensively studied in psychological and philosophical works. The most important discussion topics here are related to how media samples influence human perception and what underlying factors create that influence.

*Interestingness.* Berlyne [10, 11] identified a series of factors that influence general interest, including conflict, complexity, novelty, and uncertainty. However, these relationships are more complex, as proven in [155], as relationships may not be linear. For example, while novel information is important in inducing interest, subjects may lose interest if that information is too complex to understand. Novelty is also proposed as an important factor for interest in [169, 153]. Hidi and Anderson [87] also show that powerful emotional content has the ability to induce interest, analyzing sexual and violent content as examples. Other works look at the functional benefits brought by interest. Izard and Ackerman [95] conclude that interest is a motivational evolutionary trait, as it allows humans to explore, learn, and engage with their environment. It is presented as one of the main factors contributing to individual adaptation to the environment, survival, and development. [154, 63] also conclude that with the help of interest, in the long run, people are attracted to new possibilities and experiences. Finally, from a physiological point of view, Hess and Polt [86] show that interesting activities influence and are correlated with eye movements and pupil dilation.

*Aesthetic value.* While the aesthetic value of a picture may seem very subjective, theories suggest that some common baseline can be established that most people would agree with. Reber et al. [133] propose “goodness of form, symmetry and figure-ground contrast” as qualities necessary for an item to be deemed beautiful, as such properties would allow human assessors to process that object correctly. Fur-



thermore, with regards to visual beauty in general and to image beauty in particular, Datta et al. [38] propose that, while a normal viewer may be interested in the general effect that an image has (“how soothing a picture is to the eyes”), professional artists may be inclined to analyze other aspects, such as meaning, the use of colors and contours, sharpness and the general “rules of photography”. It is also interesting to note that in some works, interest and aesthetics have been studied as correlated concepts or, in the least, concepts that can derive from each other. This idea is best exemplified by Schmidhuber [146], who proposes that “interestingness is the first derivative of beauty: What is beautiful is not necessarily interesting. A beautiful thing is interesting only as long as it is new, that is, as long as the algorithmic regularity that makes it simple has not yet been fully assimilated by the adaptive observer who is still learning to compress the data better”.

*Memorability.* Early studies [151] regarding the memorability of images show an impressive human capacity for remembering images in the long term, even when compared with the storage capacity for other objects or concepts such as words or sentences. Furthermore, Brady et al. [13] proved that humans do not simply memorize the general scene in an image (that the authors called “gist”), but are able to encode details correctly and remember even small details and differences between images. According to [136], this capacity is further increased when subjects make a conscientious effort to memorize the images shown to them. Several other works [92, 93, 17] show memorability to be an “intrinsic propriety of images” and a dependence of memorability on the setting and objects in an image. For example, images containing people seem to be the easiest to memorize, while nature landscapes seem to be the hardest. Furthermore, time plays an important factor in memorability, whether we are talking about the difference between short-term and long-term memory, as presented in [26] or about the time a subject spends looking at an image [136].

*Violence* represents a diverse subject, given its inherent subjectivity and its perception, that can be different from society to society and from generation to generation. Ardent [5] studies violence from a modern perspective, going through some of its possible factors such as “power, strength, force, and authority”. At the same time, Galtung [66] attempts to study it from a cultural perspective, noticing the intra-cultural difference of perception of violence. While these works represent politically-oriented studies on violence, numerous other researchers studied the impact of visual violence in TV, movies, and media. Culbert [35] studies two televised violent events in 1968 (the Tet Offensive and Chicago’s DNC) and analyses the way these events changed public opinion or affected viewers at that time. The same impact is studied in [91], where the authors talk about short- and long-term effects of over-exposure to violence, including the desensitization of casual viewers and the effects on children and young adults.

*Affective value and emotions.* Psychology is the domain that started to look at the impact of emotional images on human reactions. Valdez and Mehrabian [174] explored the link between colors and the emotions that images are supposed to infer. From a different perspective, Chen and Sun [22] studied the mechanisms that allow emotions conveyed by multimedia teaching material to affect students and their learning performance. Furthermore, understanding emotions may prove useful for understanding other concepts. For example, boredom is used as an antonym of interestingness in [154, 61], and, while not precisely direct opposites, interestingness pushes human subjects towards activity and boredom pushes humans towards inactivity and limits the maximum level of interest that can be achieved [11]. Regarding the 3D representation of the VAD space, generally, valence and arousal are considered to be the most important and most frequently researched [160]; however, some scientists propose a fourth additional dimension, namely “novelty” or “unpredictability” [62], as

the addition of this dimension would better represent certain corresponding emotions from the discrete emotional space, the most relevant of them being “surprise”.

## 2.3 Datasets and user studies

Gathering an adequate dataset represents one of the most critical preliminary aspects of creating automated systems to predict such subjective proprieties. While datasets are essential in general for machine learning tasks, in this particular case, some additional matters must be taken into account, such as the difference in opinion between annotators, given the inherently subjective nature of the analyzed multimedia data. Table 2.3 summarizes the primary datasets used for predicting the concepts defined in the previous section, indicating the type of media files included in the dataset (image or video), the list of annotated concepts, and the types of annotators.

While most of the datasets are annotated by human assessors, either through crowdsourcing or through the use of “trusted” annotators that know the task well and are, in some cases, monitored continuously by super-users or master annotators, other approaches involve extracting their annotations from social media platforms directly. In this latter case, standing out are datasets that incorporate information from Flickr or Photo.net<sup>3</sup>, platforms that already provide some types of automatic or user-based annotations.

Interestingly, researchers also create an extensive collection of datasets that annotate more than one concept, an approach that may be very useful for predicting subjective concepts in the context of integrating covariates in the feature set and for analyzing inter-concept correlations. The visInterest [160] dataset, composed of 1,005 images, is collected for the study of interestingness and some of its components theorized in physiological works, such as coping potential, complexity, arousal, etc. Another example from this category is represented by two datasets created in the

---

<sup>3</sup><https://www.photo.net/>



Table 2.3: A list of relevant datasets for the subjective concepts we analyze in this thesis. We present the types of media annotated in the dataset (image or videos), the annotations provided by the authors and the types of annotators used: c - annotations collected through crowdsourcing, t - trusted annotators, w - annotations performed via social media websites, u - unknown annotation sources.

Media type	Dataset	Annotations	Annotators
image	Scene categories, interestingness [74]	interestingness	c
	Memorability, interestingness [74]	interestingness, memorability, aesthetics, unusualness, etc.	c
	visInterest [100]	interestingness, arousal, quality, coping potential, complexity, naturalness, familiarity, pleasantness	c
	LaMem [102]	37 memorability	c
	IAPS [109]	VAD space and amusement, anger, awe, fear, contentment, 37 disgust, excitement, sadness	t
	Abstract paintings [116]	amusement, anger, awe, fear, contentment, disgust, excitement, sadness	c
	Emotion6 [130]	VA, anger, disgust, fear, joy, sadness, surprise, neutral	c
	15K Flickr [144]	beauty	c
	Photo.net [38]	aesthetics, originality	w
	Aesthetics and interestingness [51]	aesthetics, social interestingness	w
	AVA [120]	aesthetics	w
image & video	MediaEval Predicting Media Interestingness [48], [47]	interestingness	t
video	Youtube dataset [96]	interestingness	t
	gifInterest [77]	interest, aesthetics, VA, curiosity	c
	NHK [167]	28 aesthetics	u
	VideoEmotion [97]	anger, anticipation, disgust, fear, joy, sadness, surprise, trust	t
	GIFGIF [98]	amusement, anger, contempt, disgust, embarrassment, fear, guilt, happiness, pleasure, etc	c
	LIRIS-ACCEDE [1]	VA, violence, fear	c
	Movie Memorability [25]	memorability	t
	Webcam	interestingness	t
	MediaEval Predicting Media Memorability [24], [31]	memorability	t
	VIF [83]	violence	c
	MediaEval Violent Scenes Detection [44], [45], [46], [159], [158]	violence	t

same work [74]. The authors considered two publicly available datasets, one on scene classification [123] composed of 2,688 images and one on memorability [94] composed of 2,222 images, and annotated them with interestingness values. Another dataset annotated with several concepts is gifInterest [77], where the authors create annotations for interestingness, aesthetics, curiosity, and the violence-arousal space. This dataset is composed of 6,119 video samples, encoded as GIFs. For the prediction of media memorability, a large image-based dataset consisting of over 58,000 samples is presented in [102]. For predicting affective content, authors take several types of approaches, given the different ways emotions are interpreted. While most of the datasets provide annotations for the VAD or VA emotional space [109], there are some examples where only categorical emotional space is used [116].

Finally, datasets such as those published during the MediaEval benchmarking competitions, listed in Table 2.3, annotated for interestingness, memorability, and violence, are also of great importance, as they provide participants with not only a dataset of media samples but also with a common evaluation framework, consisting of concept definition and use case, training/testing data splits, metrics and comparison baselines. These datasets also represent some of the largest collections available to date on their specific tasks. Creating a common evaluation framework for specific tasks can be vital for driving the development of computer vision methods forward, as it creates an accurate baseline for comparing the performance of individual methods, algorithms, and data augmentation approaches.

The first step and backbone of creating these datasets are represented by the user studies associated with them. Based on the studies and the answers returned by the annotators, researchers can create accurate ground truth data that represents the targeted concepts. From a behavioral standpoint, researchers can also deduce the way viewers interact with multimedia data and the visual cues used by annotators in making certain decisions. Some of the most important and interesting user studies consider lists of covariates for the targeted concepts and analyze positive and negative correlations between concepts. For example, Soleymani [160] studies the link between interestingness and several other concepts, concluding that *arousal* is the most important attribute for creating interest. Simultaneously, the importance of arousal is also backed up by other works, including [59, 77]. As expected, high correlation values are also found for concepts that psychological theories mention as components of interest. Some examples would include novelty [21] and complexity [160, 2]. Coping potential, another one of the theoretical indicators of interest, is also experimented from a personality perspective. More precisely, some works [160] found that coping potential may have an adverse effect of interest for subjects with high openness trait.

In contrast, complexity has a more positive effect on the same group, when compared with subjects that displayed opposing personality traits.

Other studies deal with multiple concepts. This is often the case for *interestingness*, *memorability*, *social interestingness (or popularity)* and *aesthetics*. The studies conducted by [93, 74] conclude that interestingness and memorability are negatively correlated. The studies were conducted on a set of 2222 images. Participants are asked to give their opinion on certain aspects of images (i.e., “Is this image interesting?”, “Is this image memorable?”). In the final test regarding memorability, image samples are tested, and a distinction is created between “assumed memorability” (i.e., samples that annotators think they will be able to remember) and actual memorability. Interestingly, while actual memorability is negatively correlated to interestingness, assumed memorability is positively correlated to it, suggesting that human judgment is not adequate for memorability assessment and that memory tests must be performed to ensure a correct ground truth annotation. Another relationship that is often studied is the one between aesthetics, interestingness, and popularity. Studies [90, 74, 93] show that, while visual interestingness and aesthetics are positively correlated, the same is not true for popularity. Conversely, Gygli and Soleymani [77] find a positive correlation between popularity expressed via the number of likes received by an image and visual interestingness expressed by human annotators. The annotation protocol chosen by the authors vary. For example, Hsieh et al. [90] evaluate visual interestingness on a scale of 5 options, starting from “very boring” to “very interesting”, while social interestingness is measured by social networking scores provided by the original websites where these photos are hosted.



## 2.4 Computational approaches

In recent years, computer vision algorithms started to increasingly target tasks that try to predict the affective value of multimedia data. This is an important step forward, but it requires constant collaboration between different branches of science. To understand the affective and subjective proprieties of images, computer vision scientists need to have access to large quantities of data, which would allow them to develop their methods and ensure their scalability. As we will present in the following chapter, computer vision algorithms that predict such subjective concepts are relatively new. Most of these branches started their development in the last decade, unlike more traditional <sup>15</sup> computer vision tasks such as character recognition, object detection, and image classification. This chapter presents the advances made by computer vision algorithms in predicting these affective concepts.

### 2.4.1 Interestingness

One of the first attempts at predicting image interestingness is presented in [74]. The authors used three factors in determining the interestingness score: novelty, aesthetics, and general preferences. The authors predict these sub-concepts via traditional visual features, i.e., a LOF approach for novelty prediction, aesthetic value using features proposed by [38, 116, 101], and finally, general preference determined by a set of GIST, SIFT, and color histogram descriptors in an RBF-kernel SVM. The authors point out that general preference represented the most important feature with regards to interestingness prediction. Another approach is taken by Fan et al. [58], who conclude that dataset fusion is needed in order to obtain the best results. This may indicate that, given the subjective nature of interest, a larger-than-usual number of visual samples are needed to predict interestingness correctly.

For video sample prediction, Jiang et al. [96] use a series of visual, audio, and high-level attributes. The authors find that an early fusion of visual and audio features is the optimal approach in their experiments. Jou et al. [98] perform a comparison between sentiment features and a DNN approach, based on C3D [171], finding that sentiment features perform better. Grabner et al. [73] build a system for interestingness prediction in video streams. The authors build a complexity feature based on compressed file size and a novelty feature based on LOF, achieving good results and confirming covariate-based hypothesis theorized during user studies on interestingness. An unsupervised approach is developed in [115], where images are compared via SIFT descriptors with images with comparable subjects taken from Flickr<sup>4</sup>. Here the authors base their choice of baseline Flickr images on previous findings that conclude <sup>1</sup> that Flickr users tend to curate their image collections [105].

The MediaEval Predicting Media Interestingness [48, 47] task gave the opportunity to test several systems in the same setup with regards to dataset, training / testing splits and metrics. While many systems were submitted to the benchmarking competition, some of them stand out. Liem et al [114] propose a system that, among other information, includes features that describe the presence of humans in an image, by extracting color and geometrical descriptors for the human faces from images, concluding that many times faces attract attention and interest. Shen et al. [150] use an SVM based training model that integrates deep features extracted from the AlexNet [107] DNN model. For video processing, Ben-Ahmed et al. [8] employed <sup>1</sup> deep visual and audio features based on VGG [157] and SoundNet [6], trained with a sigmoid kernel SVM. Another relevant approach is presented by Parekh et al. [125], who emulate the annotation process, by automatically developing a pairwise comparison between samples based on deep features extracted from the AlexNet DNN

---

<sup>4</sup><https://www.flickr.com/>



model. For a complete overview of the MediaEval Predicting Media Interestingness task, we refer the reader to [29].

### 2.4.2 Aesthetic value

Several papers base their approach on previous human studies on aesthetics, composition, and general photography rules. Some essential works here include [101, 38, 39, 112]. These authors designed a large set of traditional visual features centered on human perception and that are accurately able to encode some of these principles, such as depth-of-field, rule of the thirds, and “pleasant” hue combinations, object proportions, etc. These rules, taken in their entirety or just as parts of them, are still exploited in this domain [78, 85]. Regarding more modern approaches, CNN-based systems are starting to show promising results and are implemented by several authors [85]. Furthermore, a multi-patch aggregation method, based on Inception-V3 [166] models is proposed by Wang et al. [178], while Xu et al. [184] use a combination of visual features and an attention-based DNN for predicting aesthetic value.

It is important to note that aesthetics is being intensely studied alongside other concepts such as <sup>1</sup>visual interestingness and social interestingness (or social network popularity). Several authors show <sup>1</sup>a positive correlation between visual interestingness and aesthetics [51, 74, 90] <sup>1</sup>from a computer vision perspective, while social interestingness shows negative or no correlation with aesthetics. This may result from popularity having more to do with the original poster or the current news and internet trends than with the visual quality of the posted images or videos. However, Redi and Merialdo [135] use aesthetic appeal as an indicator of social interestingness using semantic and composition features on a Flickr based dataset.

### 2.4.3 Memorability

Early methods for memorability prediction [92, 93] merge human studies with computer vision methods for image classification, using conclusions drawn from the former in designing the latter. Based on the conclusion that memorability is influenced by the objects in a scene, Isola et al. [93] create a set of algorithms based on object statistics and scene descriptors and trained an SVR-based model for memorability prediction. Another significant contribution of this work is an estimator of object type importance based on memorability ground-truth value, thus creating a method for understanding why a photo is memorable. Some experiments are also directed towards increasing the memorability of an image by modifying it. Thus, style transfer models are adopted by Siarohin et al. [152], which create modified image-seed pair later scored by a selector module, that internally uses AlexNet [107] and VGG [157]. More modern approaches fully use the power of deep neural networks. For example, visual attention mechanisms and LSTM layers [89] are deployed in a ResNet-based convolutional architecture by Fajtl et al. [57]. For memorability prediction on video samples, Shekhar et al. [149] incorporate a series of deep learning, video semantics, saliency, spatio-temporal, and color features. Interestingly, fMRI data is also tested as a predictor of memorability in [80].

Recent developments are centered around the MediaEval Predicting Media Memorability task [24, 31]. Given the opportunity to test many short- and long-term memorability prediction systems in the same setting, we must address the fact that, while both editions of the task use the same dataset, data splits, and metrics, the latest edition shows significant improvements with regards to results. Thus, given the lack of additional training data, this may indicate that participants' memorability systems are objectively better. With this in mind, two systems stand out. Azcona et al. [7] employ a large set of traditional visual features, captions, and DNN-based features, trained with SVR and BRR methods, while Reboud et al. [134] combine

captions and visual information, using a large collection of training methods. Interestingly, both participants achieved top results by creating some weighted late fusion schemes that combine results extracted from lower performance systems into a supersystem with better performance.

#### 2.4.4 Violence

Many different interpretations of violence exist in datasets targeting this concept, ranging from aggression in a public environment [185] to violence in specific contexts, such as the stands of sporting events [121] or Hollywood movies and web videos [44]. As expected, the majority of approaches for predicting this concept are based on video sample assessment instead of using single image prediction, as violence is an inherently temporal concept. In this context, several works stand out. Giannakopoulos et al. [68] deploy motion-based visual features and audio features in an early fusion scheme that is trained via a kNN binary classifier. Gong et al. [71] use a semi-supervised approach to this problem, based on cross-feature learning. Starting from low-level features, the authors create candidates for violence detection and candidates for violent events, based on several labels such as “screaming”, “explosions”, “gun-shots”, etc., and combine the output of the two candidate systems. Several types of temporal integration has been tested for creating violence detection systems such as STIP and motion SIFT [121], collections of flow-vector magnitudes [83] or LSTM-based deep neural networks [81].

The 2011-2015 MediaEval Violent Scenes Detection [44, 45, 46, 159, 158] task proposes a common dataset and evaluation protocol for violence prediction methods. During this campaign, some methods stand out as outliers for their given years. For example, [145] employs a series of low-level visual features and audio features, trained in an MLP approach, and [129] use the same types of features trained by a hybrid K2 and Bayesian system. Similar to these two cases, many of the top-performing



systems employ multimodal feature fusion or multimodal training systems. Temporal aggregation or encoding of individual features or videos is achieved both by traditional methods [147] and deep learning methods based on LSTM [36, 37].

#### 2.4.5 <sup>1</sup> Affective value and emotions.

A large body of literature is dedicated to emotional content prediction. Zhao et al. [188] explore a set of high-level features based on harmony and the proportions in an image, linking the aesthetic appeal of visual samples with the emotions they convey. On the other hand, sentiment features [98] and arousal features based on color analysis [174] are used for deriving interestingness score [77]. Regarding the dimensional (VAD) emotional space, Sartori et al. [143] also investigate a series of color-based features for abstract painting emotions prediction. More modern, DNN-based approaches are also tested in both the VAD setup and for the categorical emotional space. Acar et al. [1] compare convolutional network approaches with low-level audio-visual features, obtaining better results with the neural network models. Peng et al. [130] employed a modified AlexNet architecture for the same task.

While other concepts may involve a binary prediction (i.e., interesting vs. non-interesting) or the regression equivalent or one-class regression, the problem is more complex for emotions. Research papers in this domain will have to employ either multidimensional regression, thus predicting samples with regards to the VAD space, or multi-class or multi-label classification, for the categorical emotional space. However, some works choose to take into account both approaches. For example, the authors in [119] create a set of novel audio-visual features that can successfully handle both types of tasks.

## 2.5 Applications

Great interest is shown for computer vision algorithms that can accurately predict and measure these concepts in the context of extensive image collections, where human input would be impossible to achieve due to the large amount of data that must be processed. While some systems must deal with the prediction of such concepts (for example, the detection of violent videos and images), others must create recommendation lists or proposals based on the prediction of subjective judgments, and others must modify the media samples so that values for certain concepts are maximized or minimized. Considering the impact of human cognitive processes on perception, reasoning, attention, and memorization process [55, 156] the importance of developing machine learning techniques for predicting the effect that media items have on the cognitive process.

*Image collection and video summarization.* Interest in this field is constantly growing, and web services that deal with the storage of huge amounts of personal photos, such as Google Photos<sup>5</sup> must also create automatic tools that can process the albums of users in order to create per-album or annual suggestions with regards to the best pictures in those collections. Such a feature represents one of the many ways websites can increase user engagement and loyalty. On the other hand, large videos can also be summarized in order to artificially create “trailers” or advertisements for those videos, based on several aspects. Thus, current works show a tendency of creating video summaries by measuring emotional content [183], interestingness [75, 76] and memorability [25]. Video summarization is very important for the ever-growing number of video and movie hosting services, as it would allow viewers to quickly and efficiently assess video samples and view them according to their personal preferences.

---

<sup>5</sup><https://photos.google.com/>

*Media recommendation.* While the movie recommendation literature was dominated by traditional approaches, based on past user activity and similarities between user voting preferences, some recent works are starting to incorporate visual and audio movie analysis in their algorithms [42], and make recommendations based on the audio-visual similarity of media items. Other approaches to movie recommendations also target more subjective matters, such as using image aesthetics for creating video features [40]. Perhaps a better-known application from this domain is the Flickr Interestingness API [6], which recommends multimedia items to users based on a measure of social interestingness. Aesthetics-based recommendation systems for image collections are also proposed by Schifanella et al. [144].

*Advertising systems.* Both traditional and online ads can benefit from introducing methods that can predict the positive or negative impact that ads have on viewers. Recent findings show that dissimilarities between the general viewer mood and the emotional message contained in the ads create the perception that ads are inherently “bad” or “annoying” [7]. Recent studies support these findings, as informativity and creativity, along with empathy, are considered factors for a positive response to advertising [110].

*Education.* Interest is considered to positively affect the educational process, as it would help students better process the information given to them [88]. Though this may seem obvious, some authors suggest that, in their current form, some traditional learning material and textbooks do not rely on using features that would be able to capture and hold attention [4]. Therefore, such measures of interest and other metrics related to concepts like memorability and creativity need to be introduced in the education environment. Multimedia materials could be selected based on such measures and used as learning tools, considering that interest creates motivation, willingness, and energy for learning and can guide career choices [82].

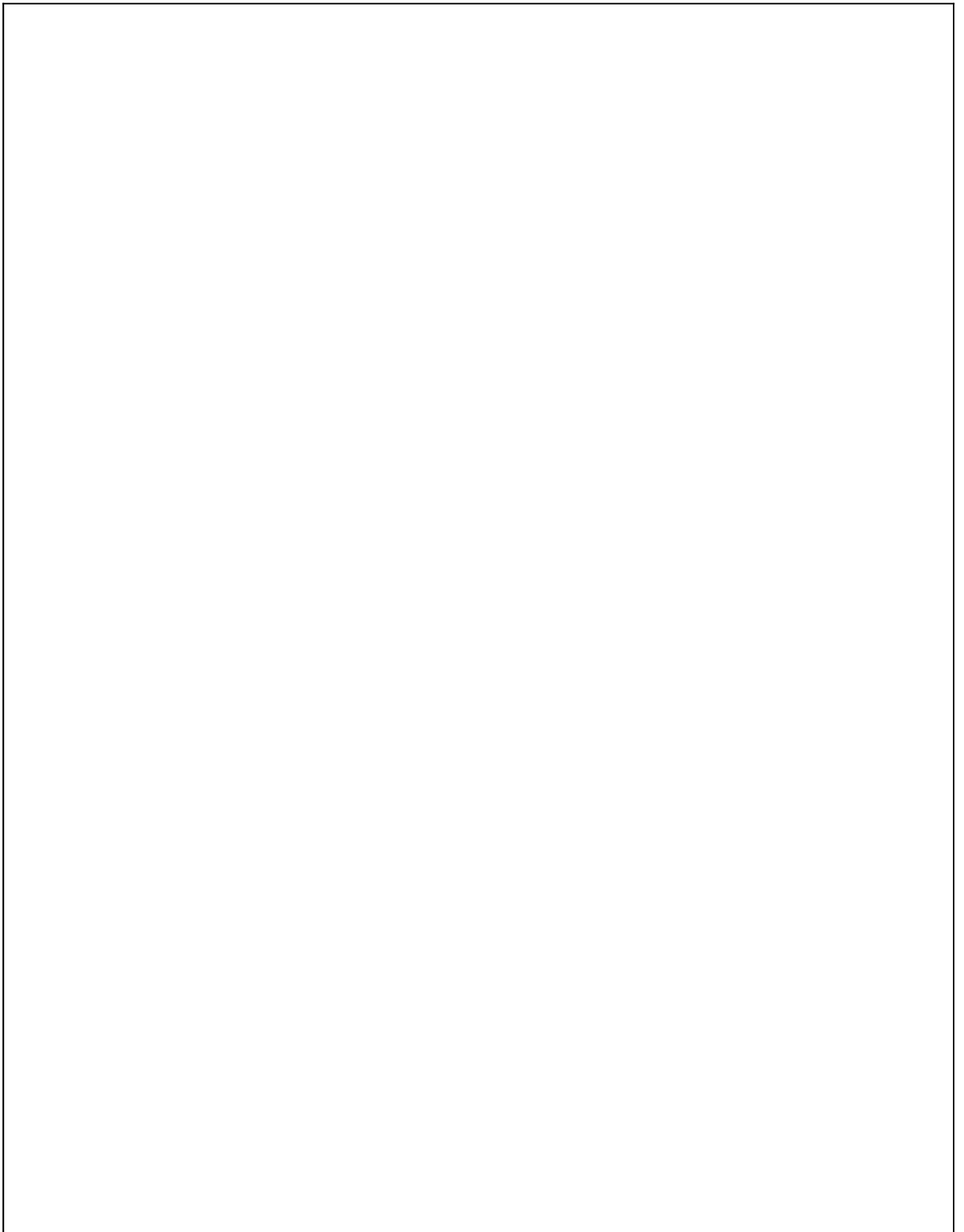
---

<sup>6</sup><https://www.flickr.com/explore/interesting/>

<sup>7</sup><http://www.tronviggroup.com/empathy-in-advertising>

## 2.6 Conclusions

In this chapter, we presented the main theoretical aspects regarding the analysis of the visual impact of multimedia data. We presented the motivations behind the need for a close collaboration between scientists from different fields of study. We have also given some examples from the <sup>62</sup>current state-of-the-art literature that show the advantages of this collaboration. We presented the definitions and some taxonomies for the concepts that will be used throughout this thesis, namely *interestingness*, *aesthetics*, *memorability*, *violence* and *affective value and emotions*. We analyzed the <sup>27</sup>state-of-the-art advances published in the current literature regarding the human understanding of these proprieties, datasets, user studies, and computational approaches. We also analyzed the subjective nature of these concepts and presented their current or future applications.





# Chapter 3

## Personal contributions

### 3.1 Datasets and evaluation

In this chapter, we will present our contribution to the creation of several publicly available datasets, including the following: (i) Interestingness10k [29]<sup>1</sup>, designed for the prediction of image and video interestingness; (ii) VSD96 [34], a video dataset for violent scenes detection; (iii) the MediaEval 2019 Predicting Media Memorability [31] a dataset composed of short videos that are annotated with short-term and long-term memorability values; and finally (iv) the MMTF-14k [41], a dataset for movie recommendation.

#### 3.1.1 Interestingness prediction

The Interestingness10k [29] dataset is a publicly available<sup>2</sup> dataset and a common evaluation framework, designed for the prediction of image and video interestingness. This dataset was tested and validated during the 2016<sup>3</sup> and 2017<sup>4</sup> editions of the MediaEval Predicting Media Interestingness tasks. My main contributions to this

---

<sup>1</sup>Paper under major review

<sup>2</sup>[https://www.interdigital.com/data\\_sets/intrestingness-dataset](https://www.interdigital.com/data_sets/intrestingness-dataset)

<sup>3</sup><http://www.multimediaeval.org/mediaeval2016/>

<sup>4</sup><http://www.multimediaeval.org/mediaeval2017/>

Table 3.1: The Interestingness10k dataset. In this table we present the composition of the image and video subsets, for both years, 2016 and 2017. Devset represents the development data, while testset represents testing data. Shown here are the number of movies, samples, average duration in seconds for the samples in the video subtask, and the number of interesting samples.

		Image		Video	
		2016	2017	2016	2017
devset	# movies	52	78	52	78
	# samples	5054	7396	5054	7396
	avg. duration (s)	-	-	1.06	1.05
	# interesting	473	714	420	646
testset	# movies	26	26 + 4	26	26 + 4
	# samples	2342	2192 + 243	2342	2192 + 243
	avg. duration (s)	-	-	1.05	2.14 + 11.4
	# interesting	241	261 + 55	226	249 + 28

dataset are represented by: (i) analyzing the overall performance of the systems submitted to the MediaEval task; (ii) analyzing the influence of features on the prediction models used during the MediaEval competition; (iii) analyzing the generalization capabilities of prediction models on our data; (iv) creating a set of recommendations with regards to system performance; (v) participating in the annotation process.

### Dataset description

This dataset is created according to a Video on Demand use case scenario, employed at Technicolor<sup>5</sup>, where participants are asked to create systems that would accurately select images or videos that would create more viewer interest in the source movie [48].

Image and video samples in this dataset are extracted from Creative Commons<sup>6</sup> licensed Hollywood-like movie trailers and segments, thus creating a publicly available set of visual data. The data is divided into image interestingness and video interestingness prediction subsets. The first step in creating this data is the extraction of video shots from complete movies, separated by camera fade-outs. While the im-

<sup>5</sup>www.technicolor.com

<sup>6</sup>https://creativecommons.org/

age subset is populated with middle keyframes extracted from those shots, the video subset is populated with the shots themselves. Some general statistics regarding this dataset, presenting both 2016 and 2017 versions, are available in Table 3.1. The dataset evolved from 5,054 devset samples extracted from 52 movies in 2016, to 7,396 extracted from 78 movies in 2017, considering that the 2017 devset data is composed of the previous year’s devset and testset combined. On the testset, in 2016, 26 movies were used, accounting for 2,342 samples, and the same number of movies was used in 2017, generating 2,192 samples. For 2017 an additional four full movies are used for enhancing the testset and test system generalization on longer excerpts (the average duration of full movie shots is 11.4 seconds, compared with 2.14 for regular samples).

Finally, with regards to system evaluation, two different metrics were chosen for the two versions of the task. For 2016, the overall MAP performance is computed, while for 2017 participant’s systems are ranked according to MAP@10.

### Overall system performance

For the overall system performance analysis, we gathered runs submitted to the MediaEval 2016 and 2017 tasks, and we analyzed the trends and improvements implemented by participants. A boxplot representation of these results is presented in Figure 3.1. In this visual representation, we also include systems developed outside of the MediaEval competition, in state-of-the-art papers, that use the same rules and validation principles as the ones used during the competition. In order to allow an overall comparison between the two years of the competition, we also provide MAP scores for the 2017 edition of the tasks. Also, for each year, three human annotator runs are represented with red dots, representing the prediction performance of humans on this dataset. While human results are above the presented systems, they still do not achieve very high results, further indicating the proposed task’s subjectivity.

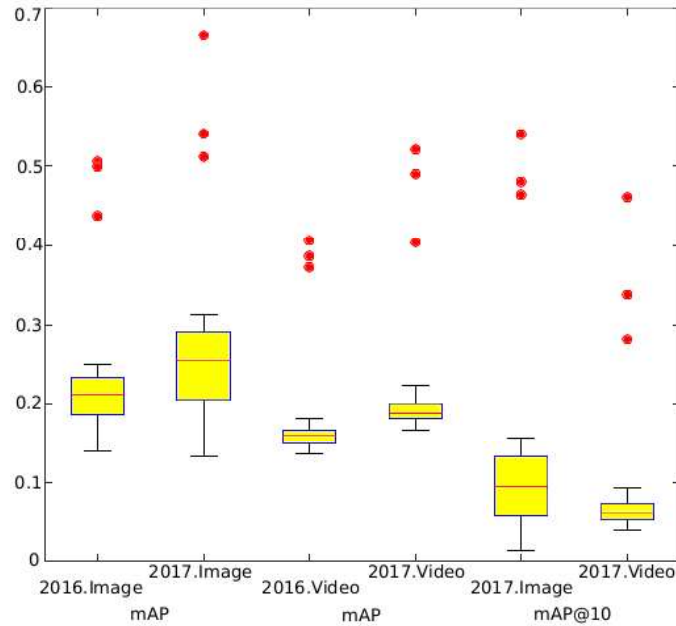


Figure 3.1: Boxplot representation of the overall system performance. Data is presented as follows: interquartile range (IQR) 50%, median values, lower and upper adjacent values calculated as  $Q1 - 1.5 \times IQR$  and  $Q3 + 1.5 \times IQR$  respectively. For reference, the performance of 3 human annotators is represented with red dots.

Regarding overall system performance, the first observation is that no systems represent either positive or negative outliers. Another clear observation is that 2017 systems performed better, with regards to MAP, than 2016 systems, indicating that a larger training set and better, more interest-oriented systems improve the overall results. In the case of the image subtask this improvement is 25.75%, from a MAP value of 0.2485 [30] to 0.3125 [125]. For the video subtask, the improvement is 22.75%, going up from a MAP value of 0.1815 [3], to 0.2228 [180].

Another critical point is that, in general, system performance for video samples is worse than that of image samples, indicating that better methods are needed for video interestingness prediction.



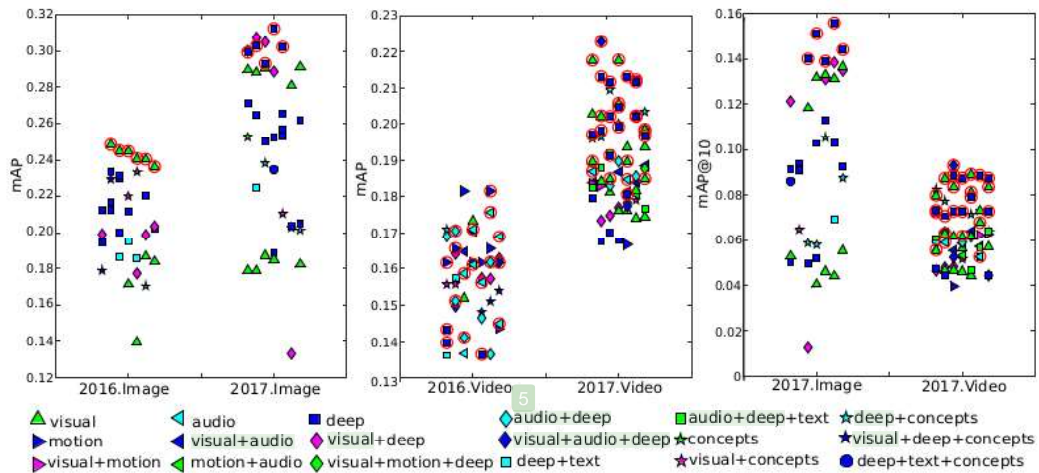


Figure 3.2: Analysis of the employed features: Year.Type represents the year of the data (2016 or 2017) and its type (Image or Video). Official metrics for 2016 data is mAP and for 2017 is mAP@10. For comparison, we also provide mAP for 2017. We represent both the participating systems from MediaEval benchmark as well as state-of-the-art approaches from literature (marked with a red circle).

### Feature-level analysis

Our analysis of the content descriptors employed by the systems submitted to this task attempts to bring to light the contributions of certain types of descriptors and, if possible, make recommendations with regards to the approaches that are best suited for interestingness prediction. We identified six main feature types that were employed by participants, as follows: traditional *visual* features, *audio*, *motion*, *deep learning*-based features, *conceptual* and *textual*. Of course, many systems use not one, but rather a combination of these types of descriptors, generating 18 employed combinations. Figure 3.2 presents the results of these approaches. We also included systems developed outside the MediaEval competition, as well as MAP performance for both years, allowing for better comparisons.

More precisely, our analysis shows that, with regards to image interestingness, systems that use deep features perform, on average, better than others, with an average MAP of 0.2297, while for video interestingness, traditional visual features



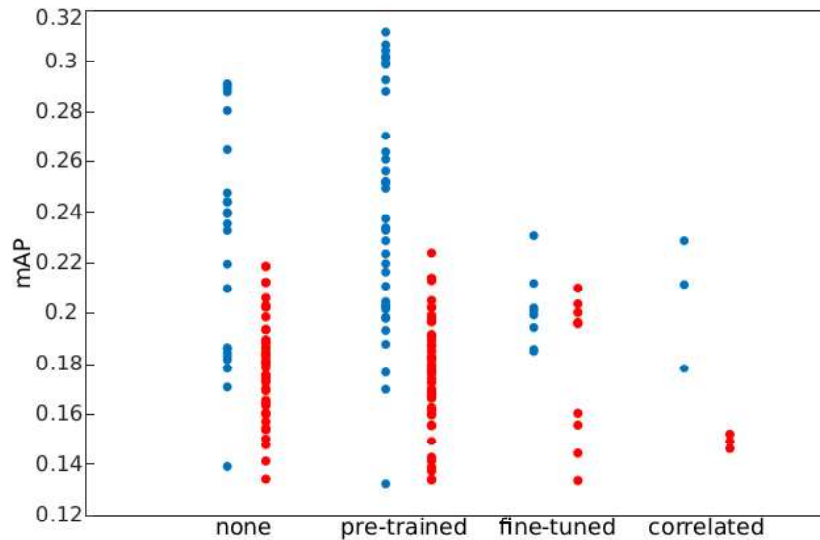


Figure 3.3: System performance for concept generalization. Blue dots represent systems used for image interestingness prediction, while red dots represent systems used for videos.

perform better, with an average MAP of 0.1798. On the other hand, when analyzing the fusion schemes employed by participants, we found that late fusion systems are the best performers on average, for both the image and video subtasks. This analysis is further presented in Table [3.2](#).

Table 3.2: Average MAP for the analyzed systems, grouped by employed features and fusion scheme.

		visual	deep	motion	audio	text	concepts	no fusion	early fusion	late fusion
Image	avg. mAP	0.2258	0.2297	-	-	0.2053	0.2157	0.2277	0.2260	0.2416
	#systems	34	53	0	0	5	11	38	35	18
Video	avg. mAP	0.1798	0.1776	0.1704	0.1746	0.1721	0.1767	0.1768	0.1731	0.1878
	#systems	49	61	14	23	6	12	51	43	30

### Generalization capabilities

We analyze three types of system generalization capabilities: (i) concept generalization, where we analyze the correlations between other concepts and interestingness, (ii) image-to-video generalization, where we test whether systems that predicting image interestingness can represent capable video interestingness predictors and finally

(iii) short-vs-long video generalization, where we compare testset performance on short and long videos.

For *concept generalization*, we theorized four types of systems, as shown in Figure 3.3. *Pre-trained* systems represent methods that are pre-trained on unrelated data, such as general image or action classification, *fine-tuned* systems represent methods that initially trained on general classification tasks and then re-trained on Interestingness10k, *correlated* systems represent methods that use data from other correlated concepts like emotional content or memorability prediction, and finally, *none* represent systems that use none of these generalization schemes, thus being trained solely on Interestingness10k data. The primary observation in this analysis is that, for the image prediction subtask, pre-trained systems significantly outperform other types of systems, with an average MAP of 0.2405, compared to 0.2208 for systems with no generalization. Unfortunately, no such statistical relevance is found for the video prediction systems. During this analysis, we identified some datasets and models that use positively or negatively correlated concepts and that are used in various system training stages. Some examples would include the methods of Shen et al. [150], that use a dataset of 0.2 million images extracted from Flickr, according to their social interestingness API<sup>7</sup> in the pre-training stage, or Erdogan et al. [56] that extracts the fully connected weights of the MemNet model [102].

With regards to *image-to-video* generalization, we analyze systems that use the same training schemes for predicting both image and video interestingness. This includes using the same set of features, training model and architectures, and pre- and post-processing methods. We also incorporate video prediction systems that employ a simple statistical approach (such as averaging) when transforming frame-level features to a global video descriptor. While only ten systems fall into this category, the correlation between image MAP performance and video MAP performance for

---

<sup>7</sup><https://www.flickr.com/explore/interesting/>

those systems, calculated via Pearson’s Correlation Coefficient, is 0.546, indicating that, although not a strict statistical proof, adapting image predictors to videos may represent a good starting point.

Finally, with regards to *short-vs-long* generalization capabilities, we separate the short testing samples from the long testing samples in the 2017 edition of the task and calculate the average MAP across these two sets. Results show an average MAP@10 of 0.0562 for short videos and 0.0751 for long videos. We attribute this to the video length difference, which may create a better differentiation between interesting and noninteresting samples.

### Recommendations with regards to system performance

Finally, we drafted a set of important observations and recommendations regarding the construction of an interestingness prediction system. These would include:

- deep features (for images) and traditional visual features (for videos) perform better than other types of descriptors;
- late fusion systems represent an obvious advantage when compared with systems that employ early or no fusion, this observation being also supported by our proposed DNN-based ensembling model;
- systems that use more than one type of classifier or regressor tend to outperform single-classifier systems;
- more modern DNN approaches, like GSM-InceptionV3 [163], can have good performances, however they do not surpass the current state-of-the-art;
- upsampling can have a positive effect on system performance, as shown in [150];
- system performance may benefit from pre-training on external data [176].

### 3.1.2 Violence prediction

The VSD96 dataset [34] is a publicly available dataset [8][9] and a common evaluation framework designed for the detection of violent scenes in Hollywood-like and YouTube [10] movies. This dataset is validated during the 2011 - 2015 editions of the MediaEval [11] Violent Scenes Detection tasks. My [27] main contributions to this dataset are as follows: (i) an overall analysis of systems that use this dataset, and (ii) an analysis of the types of features employed for violence prediction.

#### Datset description

An overview of the [29] dataset is presented in Table [3.3]. Overall, the dataset comprises annotated data from 31 full Hollywood movies, 86 YouTube videos, and 199 Hollywood-like movie clips. Several types of annotations are utilized, varying across the different editions of the MediaEval task. For 2011, 2012, and 2013, videos are segmented at shot level, via a shot boundary detector algorithm, and annotations are performed per individual shot. For 2012, 2013, and 2014, we also provide annotations at segment level, containing a starting and an ending frame number per violent segment. Finally, for 2015 annotations are done at the video clip level. Another level of annotations is represented by the definition of violence used by the annotators: (i) an objective definition, i.e., annotators are asked to determine the videos that show “physical violence or accident resulting in human injury or pain”, and (ii) a subjective definition, i.e., a video that “one would not let an 8-year old child see in a movie because it contains physical violence”. Furthermore, several metrics are used for the different versions of this dataset: (i) Cost metric for 2011, (ii) MAP@100 for 2012 and 2013, (iii) MAP2014 for 2014, and (iv) MAP for 2015.

<sup>8</sup>Data for 2011-2014 available at: [https://www.interdigital.com/data\\_sets/violent-scenes-dataset](https://www.interdigital.com/data_sets/violent-scenes-dataset)

<sup>9</sup>Data for 2015 available at: <http://liris-accede.ec-lyon.fr/>

<sup>10</sup>[www.youtube.com](http://www.youtube.com)

<sup>11</sup><http://www.multimediaeval.org/>



Table 3.3: The VSD96 dataset. We indicate the types of movie sources used (Hollywood, YouTube or Hollywood-like) year of the task (2011-2015), number of source movies, their total duration in minutes, number of segments extracted from the movies and the percentage of violent content.

Movie types	2015	2014	2013	2012	2011	# movies	duration (m)	# segm	% violence	
Hollywood movies		dev	dev	dev	dev	12	1397	21617	13.25	
				test	test	3	318	4500	19.91	
			test			3	404	6570	9.84	
						7	885	11245	12.86	
YouTube videos		test gen				86	157	86	44.47	
Hollywood-like movie clips	dev						100	1014	6144	4.42
	test						99	784	4756	4.90

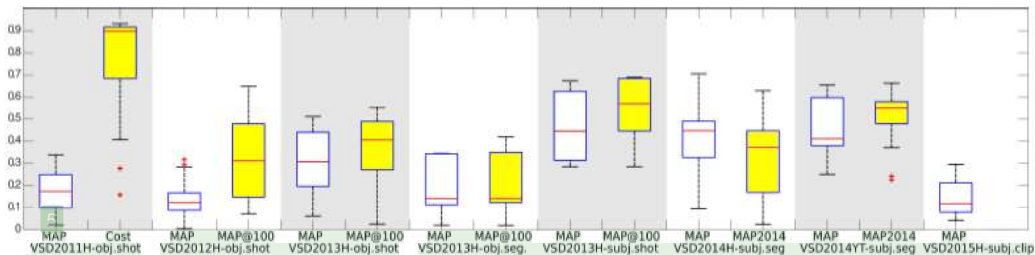


Figure 3.4: Overall performance representation. We present system performances per competition, per task, using both the original metric used during the MediaEval competition and a MAP performance, in order to allow for comparisons between editions. Boxplots are created as follows: interquartile range (IQR) 50%, median values (red line), lower and upper adjacent values calculated as  $Q1 - 1.5 \times IQR$  and  $Q3 + 1.5 \times IQR$  respectively.

### Overall system performance

Our analysis of general system performance is presented in Figure 3.4. Some improvements are evident in this analysis. For example, given the objective (denoted *obj*) definition of violence, from 2011 to 2013, MAP performance has increased, reaching 0.51 for shot-level violence prediction. An analysis based on the definition of violence can be performed, especially for the 2013 data, where both definitions are used on the same set of development and training data. Here we can notice that for systems that used shots as inputs, higher MAP and MAP@10 values are attained for the subjective definition of violence (denoted *subj*). We attribute this to better-balanced data, as there are more subjective violent samples than objective ones, i.e., 20.24%



compared with 10.49%, in the training set. Improvements have also been recorded in the 2014 version of the task, where, for segment-level prediction, a MAP value of 0.7 is attained. Also encouraging are the good results recorded on the YouTube generalization dataset (denoted *YT*). While systems for 2014 are trained on Hollywood (denoted *H*) movies, they are still capable of detecting violence in the generalization tests performed on YouTube data, indicating that systems are well trained and could perform well even in a more general understanding of violence. Also, for YouTube testing data, the class imbalance problem is significantly lower than Hollywood data, with 44.4% of this data being annotated as violent. Finally, a significant decrease in performance is registered in 2015, with a maximum MAP of 0.29. However, this may be explained as participants' systems to the 2015 task are required to predict both violence and emotional content in VA space.

### Feature-level analysis

Our analysis of the employed features shows that several types of descriptors are used in the composition of systems. These are: (i) traditional *visual* features, (ii) *audio* features, (iii) *conceptual* features, (iv) *deep* learning features. Participants also employ combinations of these four features, totaling up to 12 types of single- and multi-modality types of features. While some single modality systems, such as Dai et al. [36], achieve good results by using just traditional video features, with a MAP of 0.706 on VSD2014H-subj.seg, or Tan et al. [168] that uses an extended set of conceptual features, achieving a MAP of 0.675 and 0.674 on VSD2013H-shot.seg, multimodal systems are better performers. On average, single modality systems achieve an average MAP of 0.208, while systems that employ multiple modalities have average MAP results of 0.313, which represents a significant improvement. Furthermore, with regards to multimodal systems, four categories stand out, obtaining top results in certain subtasks: (i) *visual and audio* [72, 49], (ii) *audio and conceptual* [128], (iii)

*visual, audio and conceptual* [127, 145], and finally (iv) *visual, audio and deep* [37]. Furthermore, with regards to late fusion, ensembling systems achieve an average MAP of 0.343, thus further suggesting the advantages of late fusion schemes.

### 3.1.3 Memorability prediction

The MediaEval 2019 <sup>12</sup> Predicting Media Interestingness dataset [31], is a dataset validated during the 2019 edition of the MediaEval Benchmarking Initiative. This task requires participants to accurately predict the short- and long-term memorability for video samples. For this dataset, my main contribution is leading the organization team during the MediaEval competition.

#### Dataset description

The dataset is annotated with short- and long-term memorability ground-truth values, corresponding to human annotators' ability to remember whether they previously saw a video or not. For the short-term memorability, videos were repeated in the same annotation cycle, only tens of minutes away from their first appearance, while the long-term memorability is tested by the same annotators, 24-72 hours after the short-term cycle. <sup>25</sup> The dataset is composed of 10,000 short <sup>60</sup> soundless videos, with an average length of 7 seconds. The data is split into 80% development set data, corresponding to the videos that participants must use to develop their systems and 20% testing data. For this task, the official metric is Spearman's rank correlation.

#### MediaEval 2019 <sup>3</sup> Predicting Media Interestingness

During <sup>3</sup> this edition of the Predicting Media Interestingness task, eight teams participated in both the short- and long-term tasks. Results are encouraging, as shown in Table <sup>3.4</sup> [3.4]. The best performing systems are developed by Azcona et al. [7], with a correlation of 0.528 on the short-term prediction task and Reboud et al. [134], with a correlation of 0.277 on the long-term task. Considering the improvement in top results recorded in 2019 compared with 2018 and the fact that every system presented at this edition performs above 2018's average correlation score, we consider

<sup>12</sup><http://www.multimediaeval.org/mediaeval2019/>

Table 3.4: Results during the 2019 Predicting Media Interestingness task. For comparison, we also present the best and average results from the 2018 edition of this task.

Team	Best short-term result	Best long-term result
Insight@DCU <a href="#">7</a>	<b>0.528</b>	0.27
MeMAD <a href="#">134</a>	0.522	<b>0.277</b>
Best 2018	0.497	0.257
UPB-L2S <a href="#">32</a>	0.477	0.232
RUC <a href="#">181</a>	0.472	0.216
EssexHubTV <a href="#">111</a>	0.467	0.203
TCNJ-CS <a href="#">177</a>	0.455	0.218
HCMUS <a href="#">172</a>	0.445	0.208
GIBIS <a href="#">142</a>	0.438	0.199
Average 2018	0.359	0.173

this edition to be a success, driving forward the computational understanding of media memorability. Some general trends and processing methods that improve system performance are: using ensemble or late fusion systems, deep features, and feature dimensionality reduction.

### 3.1.4 Content recommendation

The MMTF-14K [41] is a publicly available dataset<sup>13</sup> that creates a collection of data for Hollywood movie trailer recommendation systems. While most recommender systems and datasets base their decisions on metadata-like features, consisting of user ratings, movie genres, and other related descriptors, this dataset also provides audio and visual features that can help the recommendation process, creating a multimodal decision system. My main contribution to this dataset is represented by the computation of visual deep learning-based features and visual aesthetic features.

#### Dataset description

The dataset is based on ratings and movies extracted from the popular MovieLens<sup>14</sup> dataset. User ratings are expressed on a scale of 1 to 5 stars, while the entire dataset is composed of 13,623 movie trailers, for which approximately 138 thousand users created over 12 million individual ratings. This dataset also provides metadata features that describe the movie’s genre and user-generated tags, audio features represented by BLF and i-vector features, and finally, a set of visual features<sup>50</sup> extracted from the AlexNet [107] DNN and aesthetic visual features as presented in [38, 101, 112, 78]. The MRR, MAP, and R metrics are calculated at different cutoff points (i.e., 4 and 10). These features, along with their early fusion combinations, constitute a baseline collection of methods that can be used as a baseline for comparing future methods employed by researchers who want to use this dataset.

#### Visual features

The aesthetic visual features are a collection of descriptors, aggregated by Haas et al. [78] and developed in several works on image aesthetics [38, 101, 112]. This set of 26

<sup>13</sup><https://zenodo.org/record/1225406.Xw830s8zaXw>

<sup>14</sup><https://movielens.org/>



features targets image aesthetics from three different perspectives: color-, texture- and object-based aesthetics. We present three possible feature early fusion combinations: individual features, features grouped according to the three perspectives, and a fusion scheme containing all the features. For the deep AlexNet [\[107\]](#) features, we [extract](#) [the output values of the fully connected fc7 layer](#). We provide video-level aggregation for both these features and their early fusion combinations starting from frame-level feature extraction via simple statistical aggregation, i.e., [average, median, average + variance, and median + median absolute deviation](#).

## 3.2 Predicting media interestingness

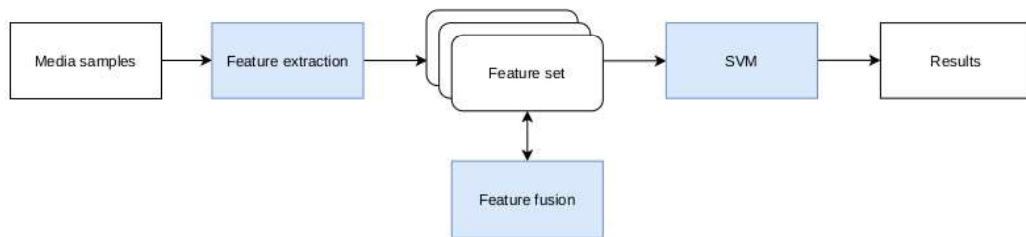
### 3.2.1 Introduction

In this chapter, we present the contributions concerning the prediction of media interestingness. We propose implementing SVM-based learning systems that use several visual features [27] as well as learning systems based on the use of aesthetic features and late fusion [28, 30]. The main contributions consist of applying a set of traditional visual features and a set of finely-grained aesthetic features to the domain of visual interestingness prediction and applying late fusion schemes in order to improve final system performance. Experiments with these approaches are carried out in the context of two consecutive benchmarking campaigns that provide two incremental datasets for both image and video interestingness prediction, namely the MediaEval 2016 [48] and 2017 [47] Predicting Media Interestingness tasks.

### 3.2.2 SVM-based learning systems

#### Motivation

As summarized in our literature survey paper [33], the concept of visual interestingness is highly subjective. Current state-of-the-art literature shows several concepts to be both positively and negatively correlated with interest. For example, while Cygli et al. [74] show valence as a positive contributor to interest, Turner and Silvia [173] show it to be negative, other examples including popularity [77, 90] and coping potential [155, 160]. While many factors can create and increase this type of subjectivity, one of the most important factors is the human annotators' personal preferences and opinions. Considering these factors, we decide to use a set of features that are traditionally used in the creation of image descriptors, thus testing a diverse baseline for this task. We present this approach in our paper [27].



<sup>33</sup> Figure 3.5: The diagram of the proposed SVM-based method. The three main stages (Feature extraction, Feature fusion and SVM) are highlighted in blue. <sup>32</sup>

### Previous work

Several works have studied the contribution of visual features to media interestingness prediction. For example, Soleymani [160] uses HOG, LBP, and GIST [123] as visual features, and trained these features using a regression that employs sparse data approximation [122] for image interestingness. Other approaches include using colorfulness [38], arousal values [116], JPEG compression size and edge distribution [101] in [74] and a semantic content detection algorithm based on Fast-RCNN [69] developed in [117]. For video interestingness prediction, Jiang et al [96] use traditional and high-level attribute features, including HSV color histogram, SIFT, GIST, Classemes [170] and style attributes [120]. Gygli and Soleymani [77] also use a set of visual descriptors, while Jou et al [98] implements a set of sentiment-based features.

### Proposed approach

Our approach consists of three phases, as described in Figure 3.5 and is described in [27]. The first stage consists of processing the media samples by employing a set of features extractors, while the second stage consists of applying feature-level fusion. Finally, the last stage consists of using an SVM-based approach for classifying the media samples.

A set of seven descriptors are extracted for each media sample. These features include: (i) color histogram calculated in the HSV space (denoted HSVHist), (ii)

dense SIFT transform with a 300 words codebook (SIFT), (iii) Local Binary Patterns (LBP), (iv) HoG descriptors calculated over densely sampled patches (HOG), (v) GIST computed with Gabor-like features (GIST), (vi) a couple of features extracted from the FC7 and Prob layers of the AlexNet architecture [107] (ANfc7 and ANprob) and (vii) the color naming histogram proposed in [175], that provides a lower-dimensional space of values for the colors in an image. For image processing, each sample is represented by this collection of feature vectors. In contrast, for video processing, we create a global video-level descriptor by averaging the vectors of all the individual frames. Regarding the feature-level fusion, we choose every combination of two individual features and, starting from that point, combinations of three best performing features, thus creating a total of 39 feature combinations for each subtask.

As previously mentioned, the final stage is represented by an SVM-based learning method. To maximize the system’s performance, we choose a broad set of experiments and start by implementing polynomial, RBF, and linear kernels. The following SVM parameters are tested for the polynomial kernels in order to optimize the results:

- polynomial degree (denoted  $d$ ) with values of 1, 2 and  $3 \times k$ , where  $k \in [1, \dots, 10]$ ;
- gamma coefficient (denoted  $\gamma$ ) with values of  $2^k$ , where  $k \in [0, \dots, 6]$ ;

while for the RBF kernels the following parameters are tested:

- cost (denoted  $c$ )
- gamma, both with values of  $2^k$ , where  $k \in [-4, \dots, 8]$ .

### Experimental setup

These methods are tested in the context of <sup>4</sup>the MediaEval 2016 Predicting Media Interestingness Task [48]. The task defines interestingness in a Video-On-Demand use case, where participants are tasked with selecting images and videos that are



most interesting for a “common viewer”. This dataset is presented and detailed in Section [3.1.1](#)

## Experimental results

The experiments are carried out in two stages. While in the initial stage, using a 10-fold cross-validation method, we select the best-performing methods with regards to MAP performance on the *devset*, in the final stage, we run the best-performing systems on the *testset*, thus obtaining the final system performance.

As a general observation, all the best-performing systems use polynomial kernel. Given the limit of five submissions per team for the final testing stage, we start by selecting the best five performers for the image and video subtasks on the *devset*, presented in Table [3.5](#). While for the image subtask, the top system with regards to the MAP metric is a polynomial SVM with  $d = 15$  and  $\gamma = 2$  that uses HSV histogram and GIST features, achieving a MAP score of 0.214, for the video subtask, a polynomial SVM represents the top system with GIST and ANprob features, and  $d = 9$  and  $\gamma = 5$ , achieving a MAP of 0.179. It is interesting to note that systems that include early feature-level fusion outperform single-feature systems. For comparison, in the image subtask, the best-performing single-feature system uses colornames, resulting in a MAP of 0.195, while for the video subtask, it is represented by a GIST-based system that achieves a MAP of 0.148. In the final stage, we use the top 2 image systems and the top 3 video systems.

Finally, the selected systems are run and tested on the *testset*. Their results are compared with the top performer and average MAP score from the MediaEval 2016 Interestingness task and presented in Table [3.6](#). As provided by the task organizers, we also present precision values for several cutoff points: 5, 10, 20, and 100. The majority of the systems we submitted present better MAP results on the *devset* they were trained on, with a single exception represented by the SIFT+ANprob



Table 3.5: Best results on *devset* for the *image* and *video* subtasks. We present the subtask (image or video), features that compose the systems, type of SVM employed and results for the Precision, Recall and MAP metrics.

Task	Feature	SVM type ( $d, \gamma$ )	Precision	Recall	MAP
image	HSVHist+GIST	poly (18, 2)	<b>0.224</b>	0.05	<b>0.214</b>
image	SIFT+GIST	poly (3, 32)	0.16	0.144	0.211
image	HSVHist+SIFT+GIST	poly (9, 2)	0.3	0.034	0.197
image	colornames+any	poly (3, 2)	0.143	0.128	0.195
image	colornames	poly (2, 8)	0.107	<b>0.517</b>	0.195
video	GIST+ANprob	poly (9, 4)	0.103	0.083	<b>0.179</b>
video	ANfc7+any	poly (3, 4)	0.099	0.095	0.172
video	SIFT+ANprob	poly (24, 64)	0.087	<b>0.192</b>	0.159
video	GIST	poly (6, 8)	<b>0.121</b>	0.116	0.148
video	SIFT	poly (3,64)	0.109	0.059	0.147

run on the video subtask. Overall, for the image subtask, the results were below the average MediaEval values, while for the video subtask, all the runs were over the average MediaEval performance, without reaching the top performance. Considering MAP, the official metric of this task, we achieve the highest performance for the submitted systems with an HSVHist + GIST combination for the image subtask ( $MAP = 0.1714$ ) and SIFT + ANProb for the video subtask ( $MAP = 0.1629$ ).

Table 3.6: Final results of the selected systems on the *testset*. The results are compared with the top performer (ME top) and average (ME avg) MAP from the MediaEval interestingness task. Results are also compared with regards to Precision metric (P) at different cutoff values (5, 10, 20, and 100).

Subtask	System	MAP	P@5	P@10	P@20	P@100
image	ME top [114]	0.2336	-	-	-	-
image	ME avg	0.2009	-	-	-	-
image	HSVHist+GIST	<b>0.1714</b>	0.1077	0.1346	0.1423	0.0869
image	SIFT+GIST	0.1398	0.0462	0.0808	0.1000	0.0862
video	ME top [3]	0.1815	-	-	-	-
video	SIFT+ANprob	<b>0.1629</b>	0.1154	0.1500	0.1192	0.0819
video	GIST+ANprob	0.1574	0.0923	0.1269	0.1212	0.0812
video	ANfc7+HSVHist	0.1572	0.1231	0.1000	0.1077	0.0815
video	ME avg	0.1572	-	-	-	-

### 3.2.3 Aesthetic features and late fusion learning systems

#### Motivation

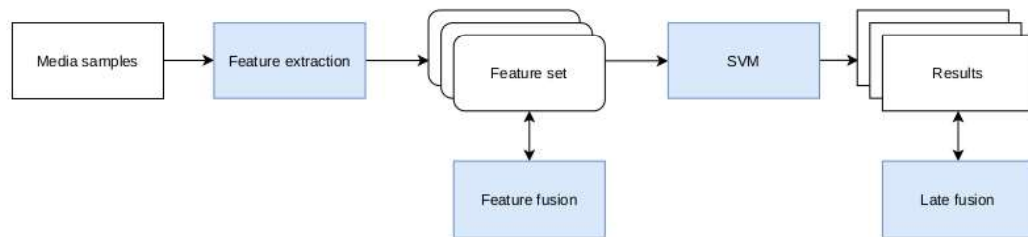
Given the previous results [27] presented at the MediaEval 2016 interestingness task, the need to implement methods that are more tuned for interestingness prediction becomes more apparent. As presented in our literature survey paper, [33], aesthetic appeal and interestingness are quite often studied together. Previous works in psychology [87] and user studies [74] found a positive correlation between aesthetics and interest. While some authors also found negative or low correlations between these two concepts [146], we nonetheless decide to extract a set of aesthetic-based features, developed in [38, 112, 101] and use these features for the prediction of media interestingness. We test this approach on the MediaEval 2016 [48] and 2017 [47] Predicting Media Interestingness Task datasets. We present these in two of our papers [30, 28].

#### Previous work

Starting from psychological works and user studies from literature, some authors have previously used aesthetic-based computer vision methods for the prediction of media interestingness. However, these approaches usually use few aesthetic cues, under the form of a low-dimensional feature vector. For example, Gygli et al. [74] used an aesthetic descriptor, composed of colorfulness values [38], arousal [116], complexity based on JPEG size and contrast and edge distribution [101]. Jou et al. [98] used simple visual features often associated with aesthetics such as brightness and balance, in creating a baseline for comparing their proposed systems.

#### Proposed approach

Our approach uses a set of aesthetic feature extractors developed in [38, 112, 101, 78], that are trained using SVM classifiers with polynomial, RBF, and linear kernels. We



<sup>33</sup> Figure 3.6: The diagram of the proposed aesthetic-based method. The four main stages (Feature extraction, Feature fusion, SVM and Late fusion) are highlighted in blue.

attempt to increase system results by employing two types of fusion experiments: early fusion and late fusion schemes. <sup>58</sup> A graphical representation of these systems is presented in Figure 3.6.

With regards to the aesthetic descriptors, three main groups of features are used in this work, as described in [78]: (i) color-based features, (ii) texture-based features, and (iii) object or segmentation-based features. Some of these are heavily inspired by research conducted in correlated domains, such as color theory, photographic practices, and image composition. The following features are part of the color-based group:

- Color values in HSV and HSL space implemented as average over the three space components (denoted HSV, HSL);
- Colorfulness implemented as quadratic-form distance and as Earth-Mover distance [101], and as standard deviation [78];
- Hue statistics, according to the findings in [112, 101] that study the importance of hues on human aesthetic perception (HueDesc);
- Hue models presence, as Li et al. [112] proposed a set of 9 hue combinations that are more pleasant (HueModel);
- Brightness, calculated as brightness contrast across the image according to the methods presented in [112];

- <sup>2</sup> Average HSV values, based on the Rule of the Thirds from image composition theory, as presented in [38] (aHSVRot);
- Average HSL values calculated around the focal point of the image, as presented in [112] (aHSLFocus).

We employ the following texture-based features:

- Edge, calculated as edge energy as presented in [112, 101] and sum of edges [78];
- Range of textures, calculated at  $3 \times 3$  bounding boxes as presented in [78] (denoted Texture);
- <sup>4</sup> Entropy of the red, green and blue spaces, as described in [78] (RGBEntropy);
- HSV Wavelet functions, <sup>4</sup> a three level Daubechies wavelet transform, implemented by [38];
- Low depth of field indicator, as presented in [38] (DoF).

Finally, the following object-based features are employed:

- Size of the largest 5 segments in an image, as proposed by [38] (denoted LargSegm);
- Centroids for the 5 larges segments in an image, as described by [38];
- HSV and brightness average values for the largest 5 segments, as proposed by [38, 112] (HueSegm, SatSegm, ValSegm, BriSegm);
- Color model for the largest 5 segments, calculated based on average color spread and complementary colors [38] (ColorSegm);
- Coordinates of the larges 3 segments, as presented by [112] (CoordSegm);
- <sup>4</sup> Mass variance and skewness for the largest 3 segments [112] (MassVarSegm, SkewSegm);
- Contrast between segments, between the HSV and blur attributes, as described in [112] (ContrastSegm).



We use the same early fusion and SVM training parameter variance schemes like those presented in Chapter 3.2.2. Furthermore, in the final step, we deploy a series of traditional statistical late fusion methods to increase system performance. In general, late fusion, or ensemble learning, is defined as a series of methods that, by combining the classification or regression outputs of several weaker learning systems, called inducers, can provide a better set of predictions for a given problem. The methods we use in this work are the following: (i) *CombSum*, (ii) *CombMin*, (iii) *CombMax*, (iv) *CombMean*. The first of these methods consists of summing the prediction outputs of the inducer systems, while *CombMin* and *CombMax* take the minimum and the maximum value respectively of the inducer’s prediction outputs. The last method consists of a weighted summing of the inducer outputs, according to the formula:

$$CombMean(Img) = \sum_{i=1}^N w_i o_i \quad (3.1)$$

where  $N$  represents the number of inducers selected for the experiment,  $o_i$  represents the outputs of individual inducers and  $w_i$  represents the weight applied to each inducer. For our experiments, we choose the following values for  $w_i$ :  $w_i = \frac{1}{2^{rank(i)}}$ . The  $rank(i)$  function returns 0 for the best performing inducer, 1 for the second best and so on.

### Experimental setup

Experiments are conducted on the datasets presented at the MediaEval 2016 and 2017 Predicting Media Interestingness Task. While the 2016 version of this dataset was also used for the experiments in Chapter 3.2.2, the 2017 version represents an extension of that dataset, with more samples for the training and testing sets. Also, for the 2016 version of the dataset, we only conducted experiments on the image subtask. Both the datasets are presented and detailed in Section 3.1.1.



Table 3.7: Final results of the systems on the MediaEval 2016 image Interestingness testset. These results are compared with the top performer (ME top) and average value (ME avg) presented during the benchmarking competition, according to the official MAP metric. Our top five systems are presented, all of them being late fusion systems, along with the best early fusion and inducer system.

Approach	MAP	Description
Late fusion	0.2485	CombMax (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2451	CombMean (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)
Late fusion	0.2448	CombMean (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2408	CombSum (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2403	CombMax (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)
Early fusion	0.2363	SatSegm + MassVarSegm + SkewSegm
ME top [114]	0.2336	
Inducer	0.2057	aHSVWavelet or SatSegm
ME avg	0.2009	

## Experimental results

Regarding the 2016 version of the dataset, the results of the experiments are presented in [30]. Similar to the experiments presented in Chapter 3.2.2 individual features performed worse than early and late fusion combinations. Table 3.7 presents the results.

It is interesting to note that, even though individual inducer systems did not outperform the MediaEval top system, represented by Liem et al [114], they did perform above average. The five best performing inducers are, in order of MAP performance: aHSVWavelet and SatSegm, both with a MAP of 0.2057 and HSV with a MAP of 0.2051. We record an increase in performance when employing early fusion schemes. In this case, early fusion results surpass the top MediaEval performance. The best early fusion schemes are as follows: SatSegm + MassVarSegm + SkewSegm

with a MAP of 0.2363, aHSVWavelet + HueSegm + SatSegm MAP performance of 0.2261 and HSL + LargSegm + BrightSegm with a MAP of 0.2232.

The late fusion systems achieve even better performances. Table 3.7 presents the top five late fusion systems. In this case, the best performing system is a CombMax late fusion scheme, applied to early fusion feature combinations of aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm, attaining a MAP score of 0.2485. Interestingly, object-based features predominantly produce better results in this experimental setup, whether they are employed in early or late fusion experiments. This observation may indicate a human annotator preference towards judging the interestingness of images based on the most salient objects in the scene.

The second part of our experiments is carried out on the MediaEval 2017 dataset, both for the image and video subtasks. The results are presented in Table 3.8

Systems developed for the video subtask perform better, having results above the average MediaEval score. The best performing system on the image subtask is a CombMean late fusion scheme that uses aHSVRot + aHSLFocus and HSV + MasVarSegm + LargSegm early fusion features ( $MAP@10 = 0.5555$ ), while on the video subtask, it is again a CombMean late fusion scheme that uses LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture early fusion features ( $MAP = 0.0732$ ). As is the case for the 2016 experiments, the late fusion systems perform better than the early fusion systems, which in turn perform better than the individual inducers. Another general observation is that the RBF kernel shows optimal results for this dataset. Surprisingly, better results are achieved for the video subtask than for the image subtask. This may result from the training phase, which is adapted to a MAP@10 setting, which perhaps does not allow for a good enough separation and therefore training between the image samples. Finally, we observe that CombMin and CombSum strategies do not improve the results of their individual inducer components.

Table 3.8: Final results of the systems submitted at the MediaEval 2017 Interestingness task. These results are compared with the top performer (ME top) and average value (ME avg) presented during the benchmarking competition, according to the official MAP@10 metric and to the additional MAP metric. For the proposed systems, results on the *devset* are also presented.

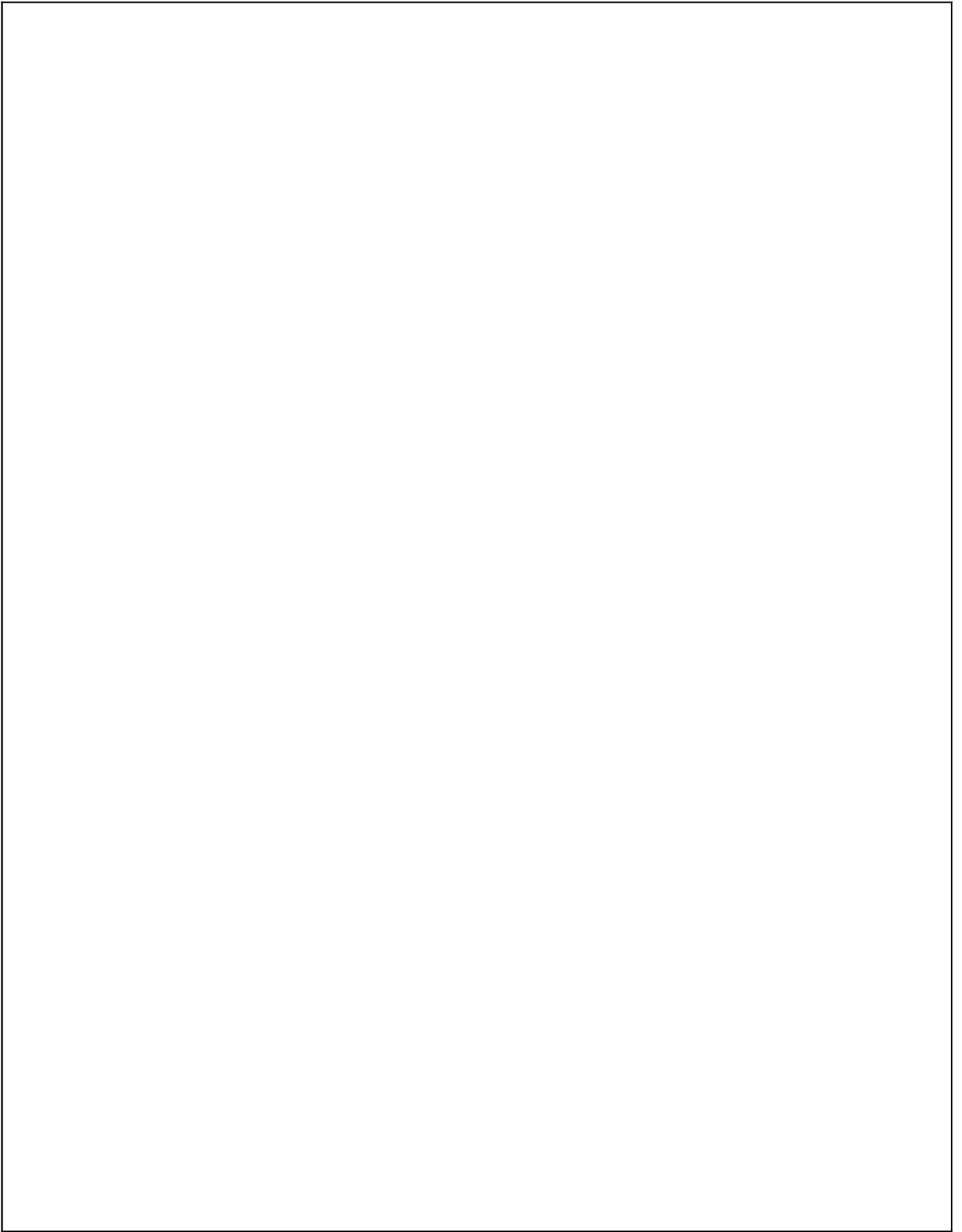
Subtask	Approach	MAP@10 devset	MAP testset	MAP@10 testset
image	ME top [131]	-	0.3075	0.1385
	ME avg	-	0.2402	0.0876
	CombMean (aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0793	0.1873	0.0555
	CombMean (HSVWavelet + aHSVWavelet + aHSLFocus and HSV + HSL + aHSLFocus and HSV + MassVarSegm)	0.0793	0.1851	0.0529
	CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0821	0.1791	0.0463
	CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus)	0.0803	0.1789	0.0442
video	ME top [8]	-	0.2094	0.0827
	CombMean(LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture)	0.0725	0.2028	0.0732
	CombMax (LargSegm + ValSegm and Texture + MassVarSegm)	0.0753	0.1937	0.0619
	CombMax (Edge + Texture and HSV + MassVarSegm)	0.0732	0.1937	0.0619
	ME avg	-	0.1845	0.0827
	CombMax (Edge + Texture and HSV + MassVarSegm and HSL + Colorfulness)	0.0723	0.1843	0.0571
	CombMax (LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture)	0.0737	0.1819	0.0564

### 3.2.4 Conclusions

In this chapter, we presented our participation [27] at the <sup>68</sup>MediaEval 2016 Predicting Media Interestingness Task [48], that uses a set of traditional visual features and SVM-learning systems, a continuation of that work [30] on the same dataset that implements a set of aesthetic-based features and late fusion schemes, and the application of our aesthetic-based system <sup>10</sup>[28] on the MediaEval 2017 Predicting Media Interestingness Task <sup>35</sup>[47]. <sup>35</sup>To the best of our knowledge, our experimental results with

aesthetic-based systems on the 2016 image subtask still <sup>35</sup> represent the current state-of-the-art, therefore proving the value of such an approach and further exploring the correlations between visual aesthetics and interestingness. Furthermore, the improvements brought by the late fusion approaches can represent an important precedent for future developments.





## 3.3 Predicting violent scenes

### 3.3.1 Introduction

<sup>43</sup>In this section, we present our contribution to the prediction of violent scenes in movies and in YouTube <sup>15</sup>surveillance videos. This approach employs a ConvLSTM [182] structure that processes visual features created by processing video frame differences with a VGG [157] network. Experiments with this approach are validated on two datasets: the MediaEval 2015 Violent Scene Detection dataset [158] and the VIF dataset [83].

### 3.3.2 Temporal deep learning systems

#### Motivation

The detection of violent scenes and events is an inherently temporal analysis; therefore, we choose to implement <sup>49</sup>state-of-the-art approaches with regards to the analysis of video sequences. While traditional methods based on motion features such as STIP and HMP [168] have been tested in literature, we wish to continue our work in analyzing top-performing systems [34], presented in Chapter 3.1.2 by adding this study that contains a state-of-the-art network.

#### Proposed approach

Our detection algorithm consists of an end-to-end temporal DNN with the ability to gather and recognize spatio-temporal information in video samples. The system does not directly use video frames as input for the processing stage, but differences between consecutive video frames, as proposed by [164]. By changing the input in this particular way, Sudhakaran et al. propose that the feature extracting networks will

---

<sup>15</sup>www.youtube.com

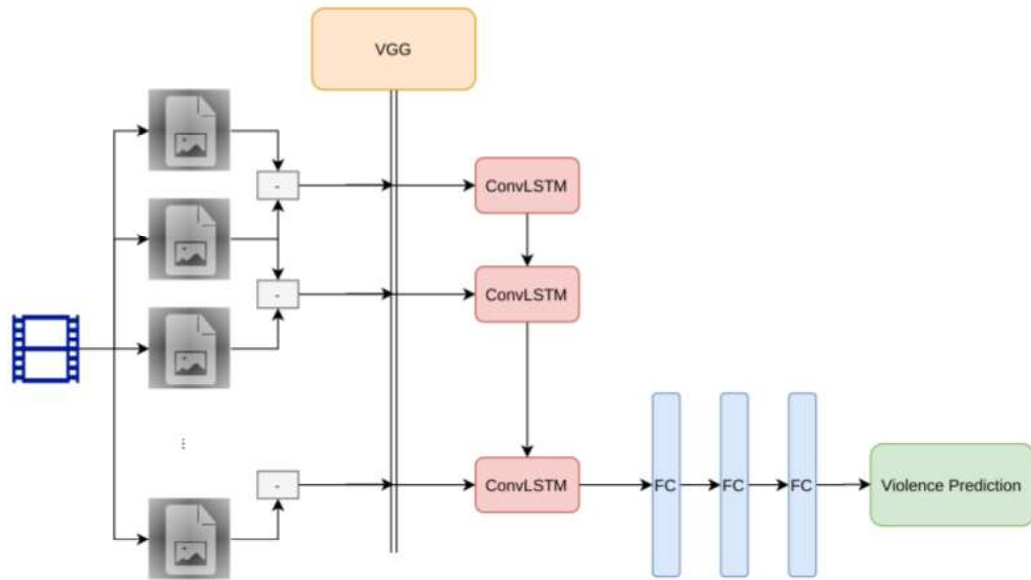


Figure 3.7: The diagram of the proposed solution. We highlight the main components, including the frame aggregator, VGG feature processor, ConvLSTM temporal aggregator and final FC layers.

be trained from the start with an internal motion correlation between its hyperparameters. The frame differences are passed after the initial stage to a VGG-19 DNN model [157], which will encode a set of features for each pair of frame differences. In the final phase, ConvLSTM [182] layers will process the <sup>36</sup>output of the VGG network. The particular setup of the ConvLSTM layer for this experiment is as follows. We use 256 filters with a dimension equal to  $3 \times 3$ , thus obtaining an output of 256 features for each processed video segment. Videos are processed with a variable-sized window of frames, equating to approximately 1 second. The final layers are <sup>57</sup>fully connected with a size of 512 neurons each, and process the ConvLSTM output in order to obtain a final decision. This network architecture is presented in Figure 3.7

## Experimental setup

Experiments are carried out on two different datasets. The first one is the MediaEval 2015 Violent Scenes Detection task [158], which contains samples extracted from Hollywood-like movies. The composition of this dataset is detailed and described in Chapter 3.1.2. The second one is the VIF dataset [83], composed of short videos extracted from YouTube. In total, the VIF dataset is composed of 267 individual video files, with a total duration of 30 minutes, split into 246 files in the training set, and 21 in the testing set. While this represents a much shorter dataset than MediaEval VSD, analyzing results in this setup will show how well the network can generalize to multiple data sources. Scenes in the VIF dataset are composed of crowd-based violence, being captured by normal security cameras, and the metric used by this dataset is Accuracy.

## Experimental results

Experimental results are presented in Table 3.9, where they are also compared with the current state-of-the-art performer on each respective dataset. The results for this approach are promising, with a maximum MAP value of 0.271 on the 2015 VSD dataset, representing a lower performance when compared with the current top result presented in [37], that achieves a MAP of 0.296, but with better results on the VIF dataset, i.e., an accuracy of 0.89, compared with the previous top results of 0.863 presented in [67]. Regarding the size of the window of frames, we tested values in {15, 20, 25, 30, 35}. While the best results, calculated on the VSD dataset are achieved for a window of size 30 (MAP = 0.271), a larger window of 35 frames also performs well, with a MAP of 0.270.



Table 3.9: Results of the proposed violence detection system, including comparison with state-of-the-art results on the respective datasets. For the MediaEval VD dataset results are presented using the official MAP metric, while for the VIF dataset results are presented according to the Accuracy metric.

Method	Window config.	VSD2015 (MAP)	VIF (Acc)
SOA	-	0.296 [37]	0.863 [67]
Our system	30	0.271	<b>0.89</b>

### 3.3.3 Conclusions

<sup>42</sup>In this chapter, we presented our approach for the task of predicting violent scenes in video samples. We developed an LSTM-based DNN, and tested the performance of this architecture on two datasets that target violence in Hollywood-like movies and surveillance videos extracted from YouTube. Results are promising, with the proposed approach performing above the current <sup>7</sup>state of the art on the surveillance dataset.

## 3.4 Predicting media memorability

### 3.4.1 Introduction

In this chapter, we present the contributions to the prediction of media memorability. Our paper [32] proposes the implementation of aesthetic and action recognition based systems to the memorability domain, and result augmentation via the implementation of a final late fusion step. My contributions to this work are represented by the implementation of the action recognition based systems and the implementation of late fusion schemes. Our approaches are tested on the publicly available dataset published during the MediaEval 2019 Predicting Media Memorability task [31].

### 3.4.2 Action-based deep learning systems

#### Motivation

In video processing, newly developed action recognition systems based on deep neural networks represent state-of-the-art approaches. These networks take advantage of temporal layers, such as LSTM layers [89], included in their architectures in order to produce better results on temporal data. Networks such as AssembleNet [139] or I3D [19] consistently represent state-of-the-art approaches at their moment of publication. We believe that using such networks would provide good results for the prediction of media memorability by accurately encoding temporal features associated with the video samples.

#### Previous work

The concept of memorability has been intensely studied by researchers in psychology, computer vision, and human studies. The memorability of an image is shown to be an intrinsic propriety of the image [92, 103]. Furthermore, Shepard [151] shows that

human visual memory has a surprisingly massive storage capacity for memorizing visual samples. From a computer vision perspective, several methods for predicting the memorability of images and videos have been tested. Generally, high-level approaches are shown to have better performance [102] with regards to memorability prediction. Good results are also achieved with some DNN-based approaches, in [60] that use an AlexNet [107] based model for creating memorable video summaries, and [25] that use several models for frame-level memorability prediction or the MemNet approach of [102].

### Proposed approach

For this approach, we use several DNN models that are pre-trained on image aesthetics and action recognition. For the aesthetic based models, a ResNet-101 architecture [84] is fine-tuned on the memorability data. At the same time, for the action recognition DNNs the TSN [179] and I3D [19] networks are used as feature extractors and augmented with the C3D features [171] provided by the task organizers. Action recognition features are passed through a dimensionality reduction step, based on PCA, and training is processed via an SVR model. A final step involves the use of late fusion schemes. The outline of this approach is presented in Figure 3.8.

The aesthetic based architecture is described in Kang et al. [99], where the authors train the ResNet-101 architecture on the AVA dataset [120] for aesthetic value prediction. For the prediction of short and long-term memorability of videos, the model is fine-tuned on the memorability dataset, using key-frames extracted in two ways: (i) from the 4th, 5th and 6th second of each video and (ii) one frame from every second in the video.

We extract the “Mixed\_5” layer and use it as a feature from the I3D model, trained on the Kinetics dataset [100], while the “Inception\_5” layer is extracted from the TSN model, trained on the UCF101 dataset [161]. We perform preliminary tests

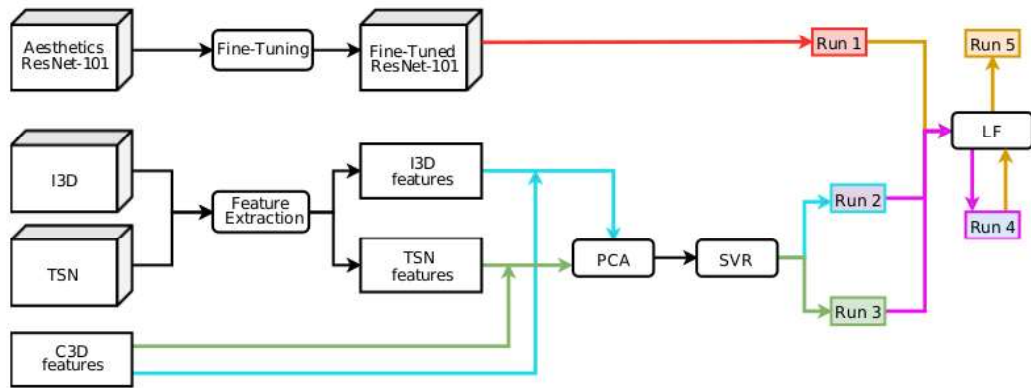


Figure 3.8: The diagram of the proposed solution. We represent the aesthetic-based network (ResNet-101) and action recognition networks (I3D, TSN, and the organizer provided C3D), their fine-tuning or extraction process and learning process, and the final late fusion (LF) stage. The components of the five individual runs submitted to the MediaEval Predicting Media Memorability task are also represented (Run 1 - Run 5).

with regards to individual I3D and TSN features, but also with regards to their early fusion combinations with the provided C3D feature. These preliminary tests favor the early fusion combinations. Finally, an SVR model is used to train these features under a randomized 4-fold data split. We tune the parameters of this SVM model using an RBF kernel with C and gamma parameters taking values of  $10^k$ , where  $k \in [-4, \dots, 4]$ .

Finally, we test three late fusion schemes that merge the action recognition systems' prediction outputs, as well as <sup>3</sup> the best action recognition system with the aesthetic-based model prediction output. The three schemes are *CombMax*, *CombMin* and *CombMean* and they are implemented in the same way as presented in Chapter [3.2.3](#)

### Experimental setup

Experiments are carried out on <sup>3</sup> the MediaEval 2019 Predicting Media Memorability task [31](#). The setup of this dataset, including the number of samples and data splits,



Table 3.10: Results of the proposed memorability systems, including preliminary tests on the devset and official results on the testset, according to the official Spearman’s  $\rho$  metric. We also include a comparison with the top and average scores registered at the MediaEval task. The five runs are denoted r1 - r5.

Run	System description	Devset - Spearman’s $\rho$		Testset - Spearman’s $\rho$	
		short-term	long-term	short-term	long-term
	ME top [7] [134]	-	-	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	0.450	0.228
	ME avg	-	-	0.448	0.206
r4	LF Action (r2 + r3)	0.466	0.200	0.439	0.218
r1	Aesthetic	0.448	0.230	0.401	0.203
r3	Action (C3D + I3D)	0.433	0.204	0.386	0.184

is presented and detailed in Chapter 3.1.3. A full comparison with results from other participants to the MediaEval benchmarking competition can also be found in that section of the thesis.

### Experimental results

The experiments are again carried out in two stages. While the first stage represents the development and validation of the systems on the *devset*, the second stage represents the deployment of the selected systems on the *testset*. Results are presented in Table 3.10 where they are also compared with the top performer and the average scores from the MediaEval task.

As previously mentioned, several variations are used in the training stage of the aesthetic DNN approach. For the short-term memorability task, two training approaches, i.e., training with the 5th frame and training with one frame per second, produce similar scores, with a Spearman  $\rho = 0.448$ , while in the long-term memorability task using the 5th frame produces better results, with a Spearman  $\rho = 0.230$ . Considering that the large majority of videos in this dataset present only one visual scene, a more extensive training dataset, as is the case for the multi-frame approach, may not be beneficial, as all frames in a video could be similar. In general, however, there is little difference between the results reported for these different approaches.

For the action recognition systems, individual systems are outperformed by early fusion schemes. Results for individual systems on the devset are as follows for the short-term memorability: TSN  $\rho = 0.418$ , I3D  $\rho = 0.401$  and C3D  $\rho = 0.3521$ . This performance further drops when the PCA processing is not implemented, therefore proving the positive influence of dimensionality reduction schemes. The top 2 performing early fusion combinations on the devset of action recognition network features are as follows: TSN + I3D, with  $\rho = 0.473$  for short-term and  $\rho = 0.259$  for long-term and C3D + I3D, with  $\rho = 0.433$  on short-term and  $\rho = 0.204$  on long-term.

We propose two late fusion combinations, namely one that would merge the two best action recognition approaches and another one that would merge both the aesthetics and the action recognition approaches. As shown in previous experiments presented in this work, CombMean and CombMax produce better results than their inducers, with CombMean being the best performing late fusion scheme.

The final results on the testset, shown in Table 3.10 show that the best performing system uses a late fusion combination of aesthetic network prediction outputs and action recognition early fusion prediction outputs. Two of our runs perform above the MediaEval average results, namely the early fusion of action features represented by the TSN and I3D and the late fusion approach that merges action and aesthetic results. For the latter, the best results are  $\rho = 0.477$  for short-term memorability and  $\rho = 0.232$  for long-term memorability.

### 3.4.3 Conclusions

In this chapter, we presented our participation at the MediaEval 2019 Predicting Media Memorability task [31], that uses aesthetics and action recognition based networks for predicting short and long term memorability for video samples. The results recorded during this competition are promising and continue to enforce the idea that

late fusion systems can be successfully applied in order to increase the results of their individual inducers significantly.

## 3.5 Late fusion with deep ensemble systems

### 3.5.1 <sup>16</sup> Introduction

In this chapter, we present the contributions to the creation of deep ensembling systems. Our works [162] and [29]<sup>16</sup> propose the creation of ensemble systems that use DNNs as the main ensembling driver. <sup>48</sup>To the best of our knowledge, this type of approach represents a novelty in the field of information fusion, where so far, DNNs have only been used as inducers for traditional fusion systems. My contribution to this work is represented by (i) the creation of two novel 2-D and 3-D input transformation schemes that allow the use of multidimensional deep neural layers, (ii) the implementation of convolutional layers in ensembling systems, (iii) and the creation of a novel DNN layer, specially designed for fusion systems, called the Cross-Space-Fusion layer. The proposed systems are tested on several publicly available datasets published as part of several MediaEval tasks, using as inducers the systems that participated at their respective tasks, as provided to us by the task organizers.

### 3.5.2 Motivation

As presented in some of the previous chapters, ensembling or late fusion systems seem to be able to significantly increase the performance of inducer algorithms for subjective tasks such as visual interestingness [3.2.3] and memorability [3.4.2] prediction. Our findings in this domain are supported by other works, where ensembles managed to achieve state-of-the-art results. Examples regarding this would include video interestingness [180], video memorability [7], and emotional content analysis [165], but also domains that do not deal with such subjective concepts, examples here including the classification of human actions in videos [163]. While these approaches do use several late fusion schemes, they do so on a lower scale, using few inducers or testing

---

<sup>16</sup>Paper under major review

their approaches on a single dataset. Furthermore, the ensembling schemes proposed by these authors are mostly represented by statistical methods, and we believe that using more inducers and employing better, more modern, ensembling schemes will significantly improve performance.

### 3.5.3 Previous work

So far, ensembling systems have employed a set of traditional methods for driving the ensemble. Some examples are already presented in this thesis, mainly statistical methods such as CombMin, CombMax, CombMean, etc. Other popular methods from the literature include boosting methods such as AdaBoost [64] and Gradient Boosting [65], bagging methods [14], methods based on random forests [15]. For a more comprehensive understanding of late fusion systems, we refer the reader to some literature survey papers that deal with this subject [70, 106, 140]. Many taxonomies of ensembles have been proposed, taking into account the main ensembling method or inducer proprieties like combination [52, 148, 108], inducer diversity [16], inducer dependency [140], and size of the ensemble [137]. However, as we previously mentioned, our approach would represent a novelty with regards to the introduction of DNN algorithms as the primary ensembling method and with regards to the number of systems employed by the ensemble.

### 3.5.4 Proposed approach

For any standard ensemble, given a set  $S$  of  $M$  samples  $s_i, i \in [1, M]$  and a set  $F$  of  $N$  classifier or regression inducer algorithms  $f_i, i \in [1, N]$ , each algorithm will produce an output for every given sample  $y_{i,j}, i \in [1, N], j \in [1, M]$ , as follows:



$$\left\{ \begin{array}{l} S = \begin{bmatrix} s_1 & s_2 & \dots & s_M \end{bmatrix} \\ F = \begin{bmatrix} f_1 & f_2 & \dots & f_N \end{bmatrix} \end{array} \right. \Rightarrow Y = \begin{bmatrix} y_{1,1} & \dots & y_{1,M} \\ \vdots & \vdots & \vdots \\ y_{N,1} & \dots & y_{N,M} \end{bmatrix} \quad (3.2)$$

Ensembling involves the creation of an algorithm  $E$ , that aggregates the outputs of inducers and learns patterns in a training set composed of individual inducer outputs and ground truth data. These patterns are then applied on the validation set, in order to produce a new output for new samples,  $e_i, i \in [1, \dots, M]$ , that represents a better-tuned output with regards to metric. The value space of  $e_i$  can differ according to the type of task the ensemble attempts to solve. For example, in regression tasks  $e_i \in [0, 1]$  or  $[-1, 1]$ , while in binary classification tasks those values can be 0 or 1. Furthermore, for multi-label or multi-class classification,  $e_i$  will be represented by a vector of values, of equal size to the number of possible classes or labels.

The proposed DeepFusion approach deploys several types of DNN that take as input the set of inducer outputs  $Y$  and produces a new set of ensembled outputs  $e$ , according to the positive and negative biases the DNN managed to learn during the training process. We thus propose to start with the creation of a baseline deep ensembling system, based on a combination of variable-sized dense layers. This baseline will then be augmented by the addition of convolutional layers, and finally, with the addition of the novel Cross-Space-Fusion (CSF) layer. While dense based networks use a 1-dimensional input for each image and video sample, convolutional and CSF layers use 2-dimensional or 3-dimensional inputs. The purpose of these layers is similar to the purpose of convolutions in image processing: we will attempt to discover and learn spatial correlations and patterns between input values that are spatially grouped together. However, such information is impossible to extract from a 1-D vec-

tor of inputs that corresponds to each sample, created by the outputs of individual inducers. We, therefore, create a set of input transformation schemes that allow us to build 2D and 3D input structures, based on the similarity degree between individual inducers, thus making possible the implementation of convolutional and CSF layers.

### Dense networks

Dense layers are known for being able to classify input data into output categories accurately, thus representing an integral part of all DNN approaches. Considering their input-agnostic nature, we theorize that building an initial baseline network that integrates several dense layers would represent a valuable starting point in creating the network. A representation of a dense ensembling architecture is presented in Figure 3.9. We choose to vary a set of parameters of these networks in order to optimize its performance with regards to the tasks being studied. The following parameters are chosen: (i) number of layers, with values of {5, 10, 15, 20, 25}; (ii) the number of neurons per layer, with values of {25, 50, 500, 1000, 2000, 5000}; and (iii) the presence or absence of batch normalization layers. We change the values of these parameters until the best results on the chosen datasets are achieved.

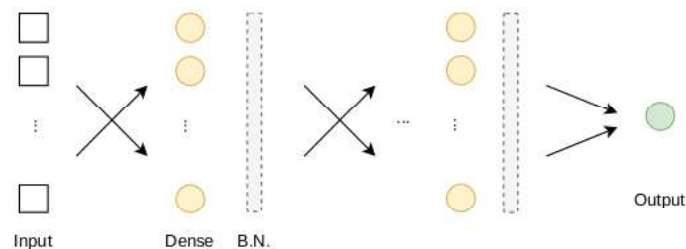


Figure 3.9: DeepFusion dense network architecture (DF-Dense): variable number of layers, number of neurons per layer and the presence or absence of Batch Normalization (BN) layers.

### Input decoration

We choose to pre-process the input data and decorate each element with output scores and data from the most similar inducers to generate spatial information. Given an image or video sample  $s_i, i \in [1, M]$ , each of the  $N$  inducer algorithms will produce a set of scores,  $Y_i$ , as described in Equation 3.3 and, as mentioned before, this kind of input has no intrinsic spatial correlation associated with it. In the first step of the input pre-processing technique, we analyze the correlation between the individual inducers  $f_i, i \in [1, N]$ . This correlation can be determined by any standard method, such as Pearson’s correlation score. However, to ensure an optimized learning process, we will use the same metric as the one the task uses as its official metric.

Given any  $f_i, i \in [1, N]$  inducer system, that produces the vector  $\bar{f}_i$  of outputs across the entire set of samples, as described in Equation 3.4 and a vector of correlation scores  $cr_i$  between this inducer and all the other inducers can be generated as presented in Equation 3.5. To generate an appropriate spatial correlation, we must use the descending ordered version of this vector, denoted  $crd_i$ :

$$Y_i = \begin{bmatrix} y_{1,i} & y_{2,i} & \dots & y_{N,i} \end{bmatrix} \quad (3.3)$$

$$\bar{f}_i = \begin{bmatrix} \bar{f}_1 & \bar{f}_2 & \dots & \bar{f}_M \end{bmatrix} \quad (3.4)$$

$$cr_i = \begin{bmatrix} cr_{1,i} & cr_{2,i} & \dots & cr_{N-1,i} \end{bmatrix} \quad (3.5)$$

As we previously mentioned, we consider both a 2D and 3D representation of the decorated input space. For the 2D representation, named *tr2D*, we apply Equation 3.6 and this input decoration scheme will be used for decorating the input for convolutional network usage. On the other hand, the two Equations presented in 3.7 describe the 3D representation, *tr3D*, with each of the two matrices being stored at different indexes in the 3rd dimension, creating a structure used by the CSF layer.

$$tr2D_{i,j} = \begin{bmatrix} \textcircled{8} & & \\ c_{1,i,j} & r_{1,i,j} & c_{2,i,j} \\ r_{4,i,j} & s_{i,j} & r_{2,i,j} \\ c_{4,i,j} & r_{3,i,j} & c_{3,i,j} \end{bmatrix}, \quad (3.6)$$

$$tr3Dc_{i,j} = \begin{bmatrix} c_{1,i,j} & c_{2,i,j} & c_{3,i,j} \\ c_{8,i,j} & s_{i,j} & c_{4,i,j} \\ c_{7,i,j} & c_{6,i,j} & c_{5,i,j} \end{bmatrix}, tr3Dr_{i,j} = \begin{bmatrix} r_{1,i,j} & r_{2,i,j} & r_{3,i,j} \\ r_{8,i,j} & 1 & r_{4,i,j} \\ r_{7,i,j} & r_{6,i,j} & r_{5,i,j} \end{bmatrix} \quad (3.7)$$

In this example, each element  $s_{i,j}$ , representing the prediction output produced by inducer  $i$  for a sample  $j$  of the input to our neural network model, is decorated with scores from similar systems,  $c_{1,i,j}$  representing the most similar system,  $c_{2,i,j}$  representing the second most similar system and so on. For the  $r$  values of our new matrix we input the correlation scores for the most similar system ( $r_{1,i,j}$ ), the second most similar ( $r_{2,i,j}$ ) and so on, with the value 1 as centroid, corresponding to the initial  $s_{i,j}$  element. The outline of the 3D decoration method is presented in Algorithm [1](#).

The spatial dimension per media sample, before the decoration process is  $N$ , in other words, equal to the number of inducer systems deployed. For the 2D approach, this dimension grows to  $(3 \times N, 3)$ , while for the 3D approach the size is  $(3 \times N, 3, 2)$ .

### Dense networks with convolutional layers

<sup>23</sup> A general presentation of the employed convolutional architecture is depicted in Figure [3.10](#). After processing the input and transforming it into a  $tr2D$  form, this input is fed into a convolutional layer. Given the  $3 \times 3$  padding of each element of the original input, we also choose to use a  $3 \times 3$  filter in our proposed architecture, therefore obtaining 10 trainable parameters in this layer. We use a stride parameter of 3, ensuring that each convolutional filter only processes similar systems. This structure <sup>47</sup> is followed by an average pooling layer that will bring the output of the convolution to



---

**Algorithm 1:** Input pre-processing algorithm for inducer  $i$ , sample  $j$

---

**Data:**  $i, j, s_{i,j}, Y_i = [y_{1,i} \ y_{2,i} \ \dots \ y_{N,i}]$ ,  $\bar{f}_i = [\bar{f}_1 \ \bar{f}_2 \ \dots \ \bar{f}_M]$   
**Result:**  $C_{i,j}, R_{i,j}$   
**begin**  
    //create the empty structures;  
     $tr3Dc_{i,j}, tr3Dr_{i,j} \leftarrow zeros(3, 3)$ ;  
    //compute the  $cr_m$  correlations;  
    **for**  $m \leftarrow 0$  **to**  $M$  **do**  
         $cr_m \leftarrow zeros(M - 1, 2)$ ;  
        //compute the  $cr_m$  correlations for each inducer;  
        **if**  $m! = i$  **then**  
             $cr_m[m, 0] \leftarrow CalcCorrelation(\bar{f}_i, \bar{f}_m)$ ;  
             $cr_m[m, 1] \leftarrow m$ ;  
        **end**  
    **end**  
    //order the inducers according to their correlation;  
     $cr_m \leftarrow Sort(cr_m)$ ;  
    //append the values to the 2-D structures according to Eq. 3.7  
    **for**  $k \leftarrow 1$  **to** 8 **do**  
         $tr3Dc_{i,j} \leftarrow AppendStructure(Y_i[cr_m[k, 1]], k)$ ;  
         $tr3Dr_{i,j} \leftarrow AppendStructure(cr_m[k, 0], k)$ ;  
    **end**  
    //append the central values;  
     $tr3Dc_{i,j} \leftarrow AppendStructure(s_{i,j}, 0)$ ;  
     $tr3Dr_{i,j} \leftarrow AppendStructure(1, 0)$ ;  
    **return**  $tr3Dc_{i,j}, tr3Dr_{i,j}$ ;  
**end**

---

the initial 1D input shape. We also test 1, 5, and 10 filters per convolution, allowing the network to perform a more extensive analysis of the similarities.

### Dense networks with Cross-Space-Fusion layers

Finally, we introduce the Cross-Space-Fusion (CSF) layer, whose general design is presented in Figure 3.11. This layer takes the 3D  $tr3D$  array and, for each group of centroids ( $tr3Dc$ ,  $tr3Dr$ ) learns a set of weights  $\alpha_{k,i}$ ,  $\beta_{k,i}$ , that process the 3D input as follows:



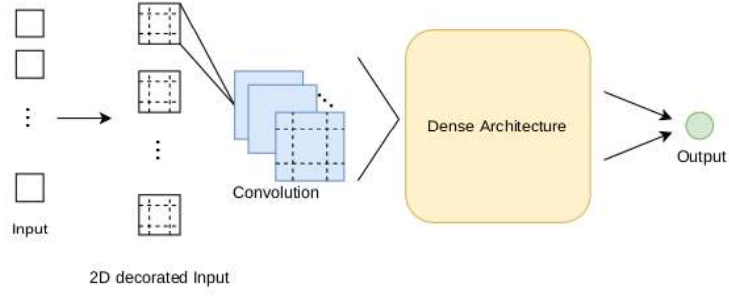


Figure 3.10: DeepFusion convolutional network architecture (DF-Conv). Represented here are the input processing stage, convolutional filters and trailing Dense Architecture.

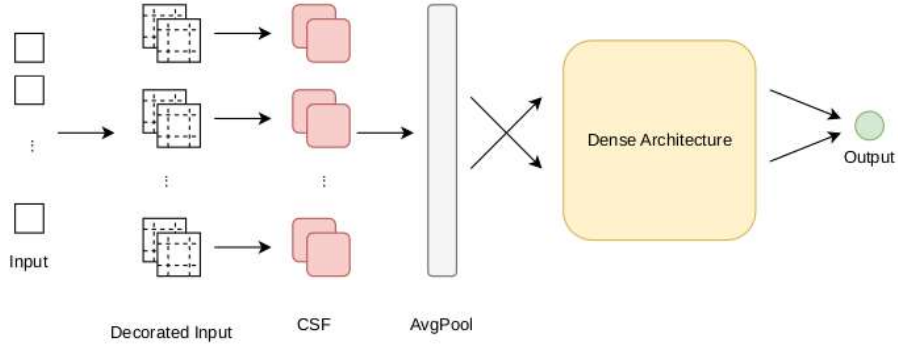


Figure 3.11: DeepFusion convolutional network architecture (DF-Conv). Represented here are the input decoration stage, CSF processing layer, Average Pooling layer and trailing Dense Architecture.

$$\begin{bmatrix} \frac{\alpha_{1,i} * s_i + \beta_{1,i} * c_{1,i} * r_{1,i}}{2} & \frac{\alpha_{2,i} * s_i + \beta_{2,i} * c_{2,i} * r_{2,i}}{2} & \frac{\alpha_{3,i} * s_i + \beta_{3,i} * c_{3,i} * r_{3,i}}{2} \\ \frac{\alpha_{8,i} * s_i + \beta_{8,i} * c_{8,i} * r_{8,i}}{2} & S_i & \frac{\alpha_{4,i} * s_i + \beta_{4,i} * c_{4,i} * r_{4,i}}{2} \\ \frac{\alpha_{7,i} * s_i + \beta_{7,i} * c_{7,i} * r_{7,i}}{2} & \frac{\alpha_{6,i} * s_i + \beta_{6,i} * c_{6,i} * r_{6,i}}{2} & \frac{\alpha_{5,i} * s_i + \beta_{5,i} * c_{7,i} * r_{5,i}}{2} \end{bmatrix} \quad (3.8)$$

The number of parameters used by the CSF layer per each centroid pair is 16, thus generating  $16 \times N$  parameters that need to be trained, where  $N$  is the total number of inducers.

Average Pooling layers finally process the output of the CSF layer, thus generating a single value for each  $(tr3Dc_i, tr3Dr_i)$  centroid group and, thus, outputting the same sized matrix as the input before the pre-processing step. We test two different setups

for data processing. In the first setup, denoted  $8S$ , all the 8-most similar inducer values are populated, while in the second setup, denoted  $4S$ , only the 4-most similar ones are populated, the rest of them being populated with zeros.

### Training protocol

We propose several essential steps in developing this late fusion approach. The first step consists of gathering all the individual  $Y_i$  vectors for each of the  $M$  samples in the training set. We then search for the best performing dense architecture by using the setup presented in “Dense networks” with regards to the number of layers, the number of neurons per layer, and the use of batch normalization. Results are tested against the validation set. The best performing dense architecture is then augmented with convolutional layers in the third step and with Cross-Space-Fusion layers in the fourth step. The input is modified for the use of the convolutional and CSF layers, as described in “Input transforms”.

For each network combination, training is performed for 50 epochs, with a batch size of 64. Loss function varied from experiment to experiment: mean squared error for the regression experiments and binary crossentropy for the classification and labeling experiments. We use an Adam optimizer [104], with an initial learning rate of 0.01.

### 3.5.5 Experimental setup

We test our proposed methods on several types of datasets: these datasets target one-class regression, multi-class regression, and multi-label prediction tasks. By testing our proposed networks on this diverse set of tasks, we wish to prove that these methods are useful in a large set of different circumstances and that they can be adapted, if needed, to a large number of use cases. We deployed our methods on the following datasets: MediaEval 2017 Predicting Media Interestingness [47] split into

an image subtask (denoted *INT2017.Image*) and a video subtask (*INT2017.Video*), MediaEval 2015 Violent Scenes Detection [158] (*VSD2015.Video*), MediaEval 2018 Predicting Emotional Impact of Movies [43] split into an arousal (*Aro2018.Video*), valence (*Val2018.Video*) and fear detection (*Fear2018.Video*), and finally the ImageCLEFmed 2019 Concept Detection [126] (*Capt2019.Image*).

While the interestingness and violence datasets are presented and analyzed in Chapters 3.1.1 and 3.1.2, respectively, and represent one-class regression tasks, the emotional impact and medical caption datasets represent new experiments. The MediaEval 2018 Emotional Impact of Movies [43] is a data set for automatic recognition of emotion in videos, in terms of valence, arousal, and fear. The data set offers annotations for two tasks, namely (i) valence and arousal prediction, a two-class regression task, consisting of 54 training/validation movies with a total duration of approx. 27 hours, and 12 testing movies with a total duration of approx. 9 hours, and (ii) fear detection, a binary classification task, consisting of 44 training/validation movies, with a total duration of more than 15 hours, and 12 testing movies with a total duration of approx. 9 hours. On the other hand, the ImageCLEFmed 2019 Concept Detection is an automatic multi-label classification image captioning and scene understanding data set [126] consisting of 56,629 training, 14,157 validation, and 10,000 test radiology examples with multiple classes (medical concepts) associated with each image. In total, there are 5,528 unique concept identifiers, whereas the distribution limits per images in the training, validation, and test sets are between 1-72, 1-77, and 1-34 concepts, respectively.

In order to create an adequate baseline of inducers, we used systems submitted to the respective tasks as inducer systems, as they have been provided to us by the task organizers. Other setups would be impractical, as they would involve training a large number of systems from the start, and considering that many times the authors of the proposed systems do not provide their source code. Furthermore, task organizers

are only able to provide us the system runs from the testset. Therefore we decided to create two types of splits on this inducer output data. In this regard, the split samples are randomized, and 100 partitions are generated to assure thorough coverage of the data, using two protocols: (i) 75% training and 25% testing (KF75), and (ii) 50% training and 50% testing (KF50). In order to avoid random splits that favor our type of approach, we generate 100 of these partitions and report the results as average values between the 100 runs. We would like to point out that, while this does not represent an accurate duplication of the original dev/test split, it does represent a disadvantage for our training stage, as we will train our deep learning fusion methods on less data than the original systems submitted to the MediaEval and ImageCLEF tasks. In this respect, we used the following number of inducers for each of the experimental datasets:

- INT2017.Image - 33 systems,
- INT2017.Video - 42 systems,
- VSD2015.Video - 48 systems,
- Aro2018.Video and Val2018.Video - 30 systems,
- Fear2018.Video - 18 systems,
- Capt2019.Image - 58 systems.

### 3.5.6 Experimental results

<sup>7</sup>In the following section, we will present the results of our experiments. For each task and set of experiments, we will provide two baselines. The first one is composed of the top-performing systems, for each dataset, recorded both during the corresponding MediaEval competition and outside of it, in <sup>55</sup>state-of-the-art works. The second set of baseline experiments are represented by a set of traditional ensembling systems that include: CombMax, CombMean, CombMean, CombAvg, CombSum, presented



Table 3.11: Final results for the convolutional architecture (DF-Conv) experiments. These results are compared with the best results from the MediaEval competition (ME top), best results from the state-of-the-art literature (SOA top), the best results from the baseline fusion systems (Emb top) and with the best results of the dense architecture experiments (DF-Dense) for the INT2017.Image task (with official MAP@10 metric), INT2017.Video (with official MAP@10 metric) and VSD2015.Video (with official MAP metric). The dataset split (dev/test, KF50 or KF75) used to produce the results is also presented.

Dataset	ME top	SOA top	Emb top		DF-Dense		DF-Conv	
Dataset split	dev/test	dev/test	KF50	KF75	KF50	KF75	KF50	KF75
INT2017.Image (MAP@10)	0.1385 [131]	0.156 [125]	0.1523	0.1674	<b>0.2316</b>	0.3355	0.2293	<b>0.3436</b>
INT2017.Video (MAP@10)	0.0827 [8]	0.093 [180]	0.0961	0.1129	0.1563	0.2677	<b>0.1692</b>	<b>0.2799</b>
VSD2015.Video (MAP)	0.296 [37]	0.303 [113]	0.3521	0.392	0.6192	0.6341	<b>0.6281</b>	<b>0.6471</b>

in previous chapters, and two boosting strategies, namely AdaBoost [64] and gradient boosting [65].

### Results for the convolutional architecture

For the convolutional architectures, we run tests on the INT2017.Image, INT2017.Video and VSD2015.Video datasets. The results are presented in Table 3.11. While the traditional early fusion schemes did improve the results, with AdaBoost being the best performer for the INT2017.Image and INT2017.Video datasets and Gradient boosting being the best performer for the VSD2015.Video dataset, their improvements are still small, especially for the KF50 setup.

On the other hand, both deep ensembling architectures significantly increase performance. The best performer in these tests is the DF-Conv architecture. As we mentioned in the description of the training protocol, the DF-Conv is built upon the best performing DF-Dense architecture, in order to analyse if the addition of convolutional layers over an already saturated dense architecture can make a difference with regards to results. Thus, the best performing DF-Dense architectures are as follows: (i) for INT2017.Image the best DF-Dense system uses 10 dense layers with 1000 neurons per layer and no BN integration, attaining MAP@10 values of 0.2316 for KF50 and 0.3355 for KF75; (ii) for INT2017.Video the best DF-Dense system has 25



layers with 2000 neurons each and BN layers, with MAP@10 performance of 0.1563 for KF50 and 0.2677 for KF75; (iii) finally, for VSD2015.Video, best performance is achieved with 5 dense layers with 500 neurons each and no BN layers, achieving a MAP score of 0.6192 for KF50 and 0.6341 for KF75.

Finally, with a single exception, namely the INT2017.Image KF50 configuration, all the DF-Conv architectures improved the results of their DF-Dense counterparts. The best performing DF-Conv architectures used 5 filters for INT2017.Image and INT2017.Video and 10 filters for VSD2015.Video. The best results for these datasets are as follows: (i) for INT2017.Image MAP@10 values of 0.2293 in a KF50 configuration and 0.3436 for KF75, (ii) for INT2017.Video MAP@10 values of 0.1692 for KF50 and 0.2799 for KF75, (iii) and finally, for VSD2015.Video, MAP values of 0.6281 and 0.6471 in KF50 and KF75 configurations respectively. These results represent a significant increase in performance, both over the ME top systems, that also represented inducers for our system, but also over state-of-the-art results, namely 120% for INT2017.Image, 200.9% for INT2017.Video and 113.5% for VSD2015.Video.

### Results for the Cross-Space-Fusion architecture

For the CSF architecture we run tests on the Aro2018.Video, Val2018.Video, Fear2018.Video and Capt2019.Image. The results of these experiments are presented in Table [3.12](#). Considering that these particular datasets and tasks are newer than the ones selected for DF-Conv architecture, no state-of-the-art systems have yet been developed for them, therefore we cannot use SOA top as a comparison baseline. Just like the case for the DF-Conv architectures, the improvements brought by the traditional fusion systems are minimal. Gradient boosting provides the best results for Val2018.Video, while AdaBoost achieves best performance for the rest of the datasets.

Table 3.12: Final results for the CSF architecture (DF-CSF) experiments. <sup>17</sup> These results are compared with the best results from the MediaEval competition (ME top), the best results from the baseline fusion systems (Emb top) and with the best results of the dense architecture experiments (DF-Dense) for the Aro2018.Video task (with official MSE and PCC metrics), Val2018.Video (with official MSE and PCC metrics), Fear2018.Video (with official IoU metric) and Capt2019.Image (with official F1 metric). The dataset split (dev/test, KF50 or KF75) used to produce the results is also presented.

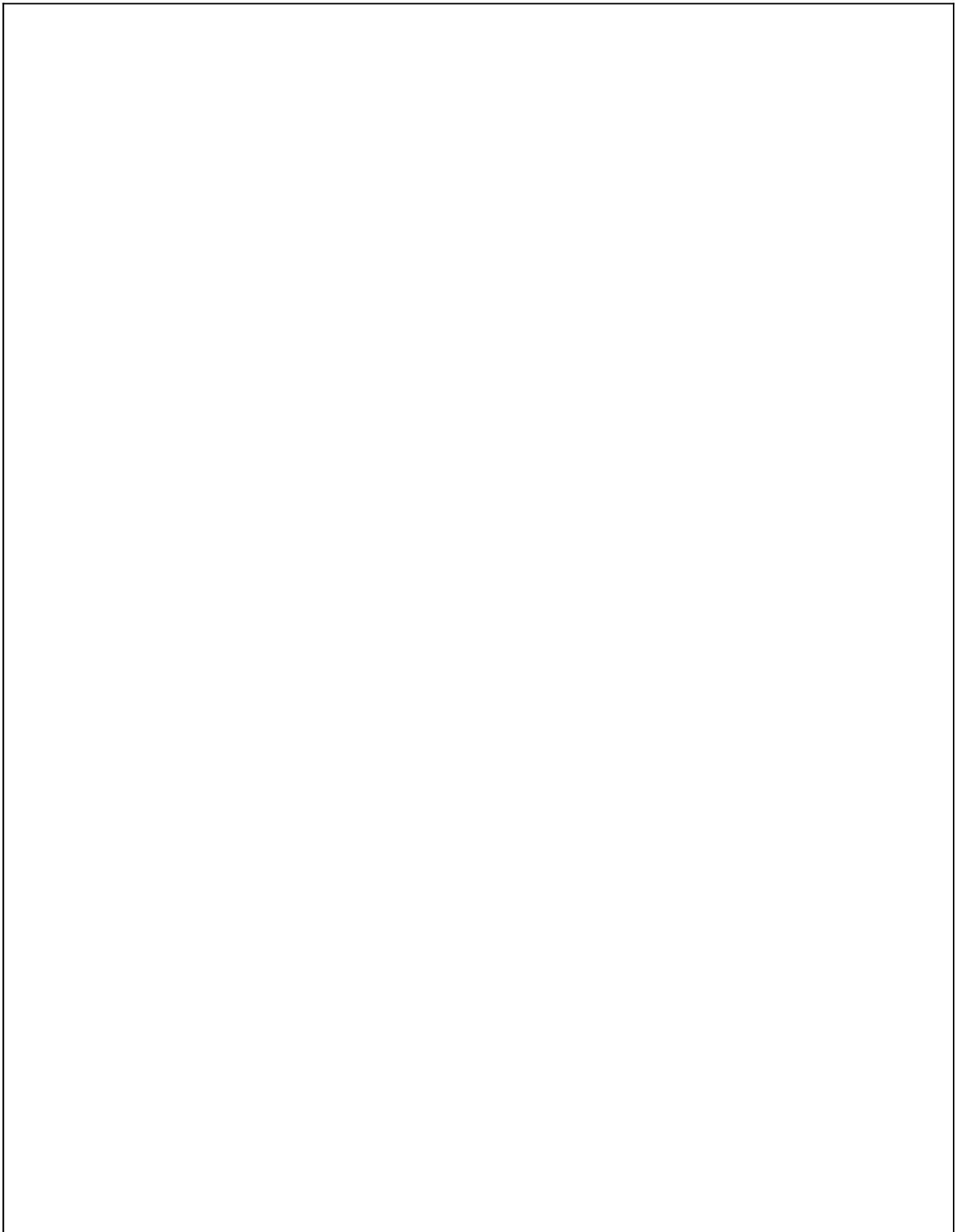
Dataset	ME top	Emb top		DF-Dense		DF-CSF	
Dataset split	dev/test	KF50	KF75	KF50	KF75	KF50	KF75
Aro2018.Video (MSE)	0.1334	0.1321	0.1253	0.0571	0.0549	<b>0.0568</b>	<b>0.0543</b>
Aro2018.Video (PCC)	0.3358	0.3547	0.3828	0.8018	0.8315	<b>0.8073</b>	<b>0.8422</b>
Val2018.Video (MSE)	0.0837	0.0814	0.0769	0.0640	0.0626	<b>0.0636</b>	<b>0.0625</b>
Val2018.Video (PCC)	0.3047	0.3372	0.3972	0.7876	0.8101	<b>0.7903</b>	<b>0.8123</b>
Fear2018.Video (IoU)	0.1575	0.1597	0.1733	0.1938	0.2129	<b>0.2091</b>	<b>0.2242</b>
Capt2019.Image (F1)	0.2823	0.2804	0.2846	<b>0.3462</b>	0.3740	0.3610	<b>0.3912</b>

Again both deep ensemble architectures significantly outperform other results. The best performing DF-Dense architectures are the as follows: (i) for both the arousal and valence datasets, we use a 5 layer architecture with 500 neurons per layer and BN layers; (ii) for Fear2018.Video the best performing architecture employs 10 dense layers with 500 neurons, without BN integration; (iii) finally, for Capt2019.Image again the best performing architecture uses 5 layers with 500 neurons and no BN.

Regarding the CSF architecture, the results are further improved when compared with the DF-Dense approach. For the arousal and valence runs, the optimal *tr3D* setup is *4S*, with only 4 similar systems used for decorating the input. MSE results are improved by 59.3% and 25.3% for the KF75 setup, with regards to MSE. However, the starting ME top results are already quite high, therefore a huge improvement, like the ones shown in the previous section are impossible. On the other hand, when looking at the PCC metric, the improvements are much larger, with 150.1% and 166.6%. For the Fear2018.Video, the DF-CSF architecture improves results by 42.3%, while for the Capt2019.Image data, improvements are at 38.6%.

### 3.5.7 Conclusions

In this chapter, we discussed the creation of a deep ensemble framework, that represents a novel research direction with regards to late fusion approaches. Our systems use dense and convolutional layers for combining inducer predictions, as well as the novel Cross-Space-Fusion layer. We also introduced two novel input transformation schemes that allow the implementation of convolutional and CSF architectures on inducer predictions. Our systems are tested on seven datasets that cover several types of machine learning tasks, including regression, binary classification, and multi-label classification, and provided significant improvements both over current state-of-the-art approaches and over traditional late fusion systems.





## Chapter 4

# General conclusions and perspectives

### 4.1 Contributions and publications

In this chapter I will summarize the main personal contributions to research papers published during my doctoral research program. These contributions are as follows:

- In (P2) I proposed the implementation of a set of traditional visual features for the prediction of media interestingness. Experimental validation is performed on the MediaEval 2016 Predicting Media Interestingness dataset.
- In (P3) and (P6) I proposed the implementation of a large set of finely-grained aesthetic features, based on color, texture, photographic and composition rules, for the prediction of media interestingness. The methods are validated both on the 2016 and on the 2017 versions of the MediaEval Predicting Media Interestingness datasets, as well as the implementation of early and late fusion schemes for performance optimization. To the best of my knowledge, the results recorded on the 2016 image subtask still represent the state-of-the-art with regards to MAP performance.

- In (P8), (P10), (P11), (P12) I proposed the implementation of visual methods for the creation of movie recommending systems. These research papers also produced the MMTF-14K dataset, where I provided a set of aesthetic and DNN-based descriptors as baselines for researchers that wish to use our dataset.
- (P13) currently represents the largest literature review on the prediction of media interestingness and its covariates. My contributions to this work are related to the study of computer vision approaches to the prediction of interestingness and its correlated concepts, the creation of a taxonomy model that studies the positive, negative and still unexplored correlations between interestingness and other subjective concepts, and, with a lower degree of involvement, the study of human understanding of interestingness.
- In (P14) I was the main organizer of the MediaEval 2019 Predicting Media Memorability task.
- In (P15) I proposed the implementation of <sup>3</sup> action recognition based DNNs for the prediction of media memorability. Results are validated on the MediaEval 2019 Predicting Media Interestingness, and early and late fusion schemes are deployed for performance optimization.
- (P18) represents a thorough analysis of the VSD96 dataset, aimed at the detection of violent video scenes. My main contributions to this work are represented by the overall analysis of the methods employed on this dataset by a large number of authors, a study of the influence of features on the prediction results and formulating some of the main conclusions with regards to the prediction of violence.
- (P19), a work currently under review, represents a thorough analysis of the Interestingness10k dataset, aimed at the prediction of image and video interestingness. My <sup>7</sup> main contributions to this paper are as follows: <sup>44</sup> the analysis of the overall performance of systems that use this dataset, an analysis of the influence of features on

the performance of systems, the study of the generalization capabilities of systems and recommendations with regards to system performance. Some shared contributions include: the study of state-of-the-art DNN approaches and interpretability of results, as well as the deployment of statistical, boosting and DNN-based late fusion systems for the improvement of the results recorded during the MediaEval 2016 and 2017 editions of the Predicting Media Interestingness task.

- (P17) represents a novel approach with regards to ensembling systems. The novelty here is represented by the introduction of DNN architectures as the main ensembling method for combining inducer prediction output. My main contributions to this paper are represented by the creation of an input decoration method, that facilitates a spatial grouping of similar inducers and by the implementation of convolutional layers for processing the decorated input. Validation is carried out on three regression tasks, namely the MediaEval 2017 image and video subtasks from the Predicting Media Interestingness task, and the 2015 MediaEval Violent Scenes Detection task, and, as results show, these methods greatly improve system performance. This work has continued, but newer results are currently unpublished. Newer results include the addition of another novel input decoration model, as well as the introduction of a novel DNN layer called Cross-Space-Fusion that is specially designed for processing ensemble data.

(P1) B. Boteanu, <sup>21</sup> **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.

(P2) <sup>3</sup> **M.G. Constantin**, B. Boteanu, B. Ionescu : LAPI at MediaEval 2016 Predicting Media Interestingness Task. In Working Notes Proceedings of the MediaEval 2016 Workshop, <sup>69</sup> CEUR-WS.org., ISSN 1613-0073. Hilversum, The Netherlands, October 20-21, 2016.

- (P3) <sup>6</sup> **M.G. Constantin**, B. Ionescu : Content Description for Predicting Image Interestingness. *IEEE International Symposium on Signals, Circuits and Systems – ISSCS*, July 13-14, Iasi, Romania, 2017.
- (P4) <sup>10</sup> 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 <sup>19</sup> *Multimedia Systems and Applications*, Eds. J. Benois-Pineau, P. Le Callet, 2017.
- (P5) B. Boteanu, **M.G. Constantin**, B. Ionescu : <sup>22</sup> *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.
- (P6) <sup>3</sup> **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.
- (P7) C.A. Mitrea, **M.G. Constantin**, L.D. Stefan, M. Ghenescu, B. Ionescu : Little-Big <sup>28</sup> *Deep Neural Networks for Embedded Video Surveillance*. *IEEE International Conference on Communications – COMM*, June 14-16, Bucharest, Romania, 2018.
- (P8) <sup>20</sup> 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*, <sup>24</sup> *June 12-15, Amsterdam, Netherlands*, 2018.
- (P9) S.V. Carata, **M.G. Constantin**, V. Ghenescu, M. Chindea, M.T. <sup>9</sup> *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.



- (P10) <sup>6</sup> 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.
- (P11) <sup>6</sup> 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.
- (P12) <sup>6</sup> 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: 3.4*).
- (P13) <sup>3</sup> **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: 6.131*).
- (P14) **M.G. Constantin**, B. Ionescu, <sup>3</sup> 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.
- (P15) **M.G. Constantin**, C. Kang, G. Dinu, F. Dufaux, G. Valenzise, B. <sup>9</sup> Ionescu : Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In Working Notes <sup>39</sup> Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, October 27-29, 2019.



- (P16) B. Ionescu, H. Müller, R. Péteri, D.-T. Dang-Nguyen, ... , M. Dogariu, L.-D. Ştefan, **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](https://doi.org/10.1007/978-3-030-45442-5_69), ECIR 2020 Proceedings, April 14-17, Lisbon, Portugal, 2020.
- (P17) L.-D. Ştefan, **M.G. Constantin**, B. Ionescu : System Fusion with Deep Ensembles. <sup>21</sup> ACM International Conference on Multimedia Retrieval - ICMR, October 26-29, Dublin, Ireland, 2020.
- (P18) <sup>46</sup> **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: 6.28*).
- (P19) *Paper under major review* : **M.G. Constantin**, L.-D. Ştefan, B. Ionescu, <sup>67</sup> N.Q.K. Duong, C.-H. Demarty, M. Sjöberg : Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision.



Mihai Gabriel Constantin

Universitatea Politehnica Bucuresti

Verified email at imag.pub.ro - Homepage

Multimedia Processing Machine Learning Neural Networks Deep Learning

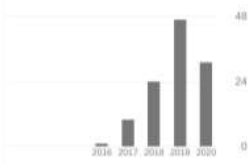
FOLLOWING

TITLE	CITED BY	YEAR
Audio-visual Encoding of Multimedia Content for Enhancing Movie Recommendations Y Deldjoo, MG Constantin, E Hamid, M Schedl, B Ionescu, P Cremonesi Proceedings of the 12th ACM Conference on Recommender Systems, ACM	25	2018
Movie Genome: Alleviating New Item Cold Start in Movie Recommendation Y Deldjoo, M Ferrari Dacrema, MG Constantin, H Eghbal-zadeh, ... User Modeling and User-Adapted Interaction, 1-63	23	2019
MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval Y Deldjoo, MG Constantin, B Ionescu, M Schedl, P Cremonesi Proceedings of the 9th ACM Multimedia Systems Conference, 450-455	16	2018
Predicting interestingness of visual content CH Demary, M Sjöberg, MG Constantin, NQK Duong, B Ionescu, TT Do, ... Visual Content Indexing and Retrieval with Psycho-Visual Models, 233-265	12	2017
LAPI at MediaEval 2016 Predicting Media Interestingness Task MG Constantin, B Boteanu, B Ionescu	11	2016
Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates MG Constantin, M Redl, G Zen, B Ionescu ACM Computing Surveys 52 (2), 25:1–25:37	7	2019
LAPI@ 2016 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversification Perspective. B Boteanu, MG Constantin, B Ionescu MediaEval	6	2016
The MediaEval 2018 Movie Recommendation Task: Recommending Movies Using Content. Y Deldjoo, MG Constantin, A Dritsas, B Ionescu, M Schedl MediaEval	4	2018
Content Description for Predicting Image Interestingness MG Constantin, B Ionescu International Symposium on Signals, Circuits and Systems - ISSCS 2017	4	2017
Predicting Media Memorability Task at MediaEval 2019 MG Constantin, B Ionescu, CH Demary, NQK Duong, X Alameddine Pineda, ... Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France	3	2019
Affect in Multimedia: Benchmarking Violent Scenes Detection MG Constantin, LD Stefan, B Ionescu, CH Demary, M Sjöberg, M Schedl, ... IEEE Transactions on Affective Computing	2	2020
Little-big deep neural networks for embedded video surveillance CA Mitrea, MG Constantin, LD Stefan, M Ghenea, B Ionescu 2018 International Conference on Communications (COMM), 493-496	1	2018
System Fusion with Deep Ensembles LD Stefan, MG Constantin, B Ionescu Proceedings of the 2020 International Conference on Multimedia Retrieval ...		2020
ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications B Ionescu, H Müller, R Peter, DT Dang-Nguyen, L Zhou, L Piras, ... European Conference on Information Retrieval, 533-541		2020
Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability MG Constantin, C Xiang, G Dinu, F Dufaux, G Valeriani, B Ionescu Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France		2019
Innovative Multi Pcnr Based Network for Green Area Monitoring-Identification and Description of Nearly Indistinguishable Areas-in Hyperspectral Satellite Images SV Cerata, MG Constantin, V Ghenea, M Chindia, M Ghenea IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium ...		2018
ISM 2019 S Alsaedi, PK Atrey, W Baker, S Bba, M Berndorf, J Benois-Frenau, ...		

Articles 1-17 SHOW MORE

Cited by

	All	Since 2015
Citations	114	114
h-index	6	6
i10-index	5	5



Co-authors

- Bogdan Ionescu  
University Politehnica of Bucharest
- Markus Schedl  
Professor at Johannes Kepler U...
- Yashar Deldjoo  
Assistant Professor, Polytechnic ...
- Paolo Cremonesi  
Dept. Computer Science, Politec...
- Hamid Eghbal-zadeh  
Institute of Computational Peirc...
- Bogdan Andrei Boteanu  
University Politehnica of Bucharest
- Claire-Hélène Demary  
Senior Scientist, InterDigital
- Mats Sjöberg  
Machine learning specialist at CSC
- Ngoc Q. K. Duong  
Senior Research Scientist, Inter...
- Stefano Cereda  
Politecnico di Milano
- Maurizio Ferrari Dacrema  
PHD Student, DEIB, Politecnico ...
- Liviu-Daniel Stefan  
University Politehnica of Bucharest
- Thanh-Toan Do  
Lecturer (Assistant Professor), U...
- Miriam Redl  
Wikimedia Foundation
- Gloria Zen  
University of Trento
- Xavier Alameda-Pineda  
Research Scientist, Perception T...
- Guillaume Gravier  
CNRS - IRISA
- Catalin Mitrea  
University Politehnica of Bucharest
- Serban Vasile Carata
- Frederic Dufaux  
Université Paris-Saclay, CNRS, ...

## 4.2 Conclusions

<sup>23</sup>This thesis presents my personal contributions to the automatic analysis of the visual impact of multimedia data, with an accent on the study of *interestingness*, *aesthetics*, *memorability*, *violence* and *affective value and emotions*. Chapter <sup>2</sup> presents an analysis of the <sup>51</sup>current state-of-the-art with regards to concept taxonomy and definitions, theories on the human understanding of subjective multimedia proprieties, datasets and user studies, computational approaches, and current applications and future perspectives on the use of these proprieties. Chapter <sup>3</sup> presents my contributions to this field. The first part of this chapter covers the datasets and benchmarking initiatives I have contributed to. Following this, the thesis presents several computer vision methods developed during my doctoral program and analyses the contributions to the current computational landscape brought by these methods. Methods presented here are related to: (i) the prediction of media interestingness via traditional visual features in an SVM learning setting, and the implementation of aesthetic-based features and statistical late fusion schemes for interestingness prediction; (ii) the detection of violent scenes via the implementation of a ConvLSTM approach; (iii) the prediction of media memorability with the help of action recognition deep neural networks; (iv) the creation of a novel deep learning based approach to ensemble learning, the creation of new input decoration methods that would allow the processing of correlated inducers in our deep fusion systems and a novel type of deep neural network layer, the Cross-Fusion-Layer, specially designed for the processing of ensemble systems.

The results presented in this thesis are promising, especially considering that the proposed deep fusion systems significantly increase state-of-the-art performance. While in general late fusion systems do require more processing power, given that the data is processed by multiple inducers, one must consider that these types of approaches will prove to be useful, given the constant improvement of GPU processing power and the advent of online services dedicated to processing massive amounts of

data in a reasonable time. I consider that such systems can be deployed in many use cases, mainly in scenarios where the results of individual systems are not good enough for a final market-ready solution, or in critical infrastructure systems, where accurate results are more important than the cost of a system.

### 4.3 Future perspectives

In continuation of this work, the most important aspect would be the implementation of systems that are better tuned for their respective tasks. Some examples are already presented in this thesis, i.e., aesthetic-based features, but I consider that, by implementing more of these types of systems based on previous research from the fields of psychology and behaviour analysis, better architectures can be constructed and their results would better benefit the multimedia community.

Furthermore, given the results of the deep ensemble system, I consider that it represents a very interesting research direction for the future. While this approach represents, <sup>17</sup> to the best of my knowledge, the first attempt in creating such deep fusion systems, future developments may include: the creation of novel input decoration methods, the addition of novel layers and training schemes for the existing layers, and studies with regards to optimizing the collection of employed inducers.





## Bibliography

- [1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos. In *IEEE International Workshop on Content-Based Multimedia Indexing*, pages 1–6. IEEE, 2015.
- [2] Peter P Aitken. Judgments of pleasingness and interestingness as functions of visual complexity. *Journal of Experimental Psychology*, 103(2):240–244, 1974.
- [3] Jurandy Almeida. Unifesp at mediaeval 2016: Predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [4] Richard Chase Anderson. Interestingness of children’s reading material. *Center for the Study of Reading Technical Report; no. 323*, 1984.
- [5] Hannah Arendt. *On violence*. Houghton Mifflin Harcourt, 1970.
- [6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.
- [7] David Azcona, Enric Moreu, Feiyan Hu, Tomás Ward, and Alan Smeaton. Predicting media memorability using ensemble models. In *Working Notes Proceedings of the MediaEval 2019.*, 2019.
- [8] Olfa Ben-Ahmed, Jonas Wacker, Alessandro Galallo, and Benoit Huet. Eu-recom@ mediaeval 2017: Media genre inference for predicting media interestingness. In *Working Notes Proceedings of the MediaEval 2017 Workshop.*, 2017.
- [9] Daniel E Berlyne. ‘interest’ as a psychological concept. *British journal of psychology. General section*, 39(4):184–195, 1949.
- [10] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.
- [11] Daniel E Berlyne. Novelty, complexity, and hedonic value. *Perception & Psychophysics*, 8(5):279–286, 1970.
- [12] Daniel E Berlyne. *Aesthetics and psychobiology*, volume 336. JSTOR, 1971.

- [13] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.
- [14] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [16] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.
- [17] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.
- [18] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.
- [19] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [20] Christel Chamaret, Claire-Helene Demarty, Vincent Demoulin, and Gwenaelle Marquant. Experiencing the interestingness concept within and between pictures. *Electronic Imaging*, 2016(16):1–12, 2016.
- [21] Ang Chen, Paul W Darst, and Robert P Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3):383–400, 2001.
- [22] Chih-Ming Chen and Ying-Chun Sun. Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. *Computers & Education*, 59(4):1273–1285, 2012.
- [23] Sharon Lynn Chu, Elena Fedorovskaya, Francis Quek, and Jeffrey Snyder. The effect of familiarity on perceived interestingness of images. In *IS&T/SPIE Electronic Imaging*, volume 8651, pages 86511C–86511C. International Society for Optics and Photonics, 2013.
- [24] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability task. *Working Notes Proceedings of the MediaEval 2018 Workshop.*, 2018.
- [25] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. Transfer learning for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*, 2018.
- [26] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2531–2540, 2019.

- [27] Mihai Gabriel Constantin, Bogdan Boteanu, and Bogdan Ionescu. Lapi at mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [28] Mihai Gabriel Constantin, Bogdan Andrei Boteanu, and Bogdan Ionescu. Lapi at mediaeval 2017-predicting media interestingness. In *Working Notes Proceedings of the MediaEval 2017 Workshop.*, 2017.
- [29] Mihai Gabriel Constantin, Liviu-Daniel Ștefan, Bogdan Ionescu, Ngoc Q. K. Duong, Claire-Hélène Demarty, and Mats Sjöberg. Visual interestingness prediction: A benchmark framework and literature review. *Under Review at: International Journal of Computer Vision*, 2020.
- [30] Mihai Gabriel Constantin and Bogdan Ionescu. Content description for predicting image interestingness. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017.
- [31] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc QK Duong, Xavier Alameda-Pineda, and Mats Sjöberg. Predicting media memorability task at mediaeval 2019. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [32] Mihai Gabriel Constantin, Chen Kang, Gabriela Dinu, Frédéric Dufaux, Giuseppe Valenzise, and Bogdan Ionescu. Using aesthetics and action recognition-based networks for the prediction of media memorability. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [33] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)*, 52(2):1–37, 2019.
- [34] Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Hélène Demarty, Mats Sjöberg, Markus Schedl, and Guillaume Gravier. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, 2020.
- [35] David Culbert. Television’s visual impact on decision-making in the usa, 1968: The tet offensive and chicago’s democratic national convention. *Journal of Contemporary History*, 33(3):419–449, 1998.
- [36] Qi Dai, Zuxuan Wu, Yu-Gang Jiang, Xiangyang Xue, and Jinhui Tang. Fudan-just at mediaeval 2014: Violent scenes detection using deep neural networks. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*, 2014.
- [37] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *Working Notes Proceedings of the MediaEval 2015.*, 2015.



- [38] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006.
- [39] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *IEEE International Conference on Image Processing*, pages 105–108. IEEE, 2008.
- [40] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 455–459, 2018.
- [41] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. Mmtf-14k: a multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 450–455, 2018.
- [42] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. Using visual features based on mpeg-7 and deep learning for movie recommendation. *International journal of multimedia information retrieval*, 7(4):207–219, 2018.
- [43] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, and Mats Sjöberg. The mediaeval 2018 emotional impact of movies task. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*, 2018.
- [44] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The mediaeval 2011 affect task: Violent scene detection in hollywood movies. In *Working Notes Proceedings of the MediaEval 2011 Workshop.*, 2011.
- [45] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The mediaeval 2012 affect task: Violent scene detection. In *Working Notes Proceedings of the MediaEval 2012 Workshop.*, 2012.
- [46] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang, and Yu-Gang Jiang. The mediaeval 2013 affect task: violent scenes detection. In *Working Notes Proceedings of the MediaEval 2013 Workshop.*, 2013.
- [47] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Duong. Mediaeval 2017 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2017 Workshop.*, 2017.
- [48] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc QK Duong, and Frédéric Lefebvre. Mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.

- [49] Nadia Derbas, Bahjat Safadi, and Georges Quénot. LIG at mediaeval 2013 affect task: Use of a generic method and joint audio-visual words. In *Proceedings of the MediaEval 2013 Workshop*, 2013.
- [50] Arturo Deza and Devi Parikh. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1826. IEEE, 2015.
- [51] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664. IEEE, 2011.
- [52] Robert PW Duin. The combining classifier: to train or not to train? In *Object recognition supported by user interaction for service robots*, volume 2, pages 765–770. IEEE, 2002.
- [53] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [54] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008.
- [55] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595, 2003.
- [56] Goksu Erdogan, Aykut Erdem, and Erkut Erdem. HUCVL at mediaeval 2016: Predicting interesting key frames with deep models. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [57] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372, 2018.
- [58] Shaojing Fan, Tian-Tsong Ng, Bryan L Koenig, Ming Jiang, and Qi Zhao. A paradigm for building generalized models of human image perception through data fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5762–5771. IEEE, 2016.
- [59] Kirill Fayn, Carolyn MacCann, Niko Tiliopoulos, and Paul J Silvia. Aesthetic emotions and aesthetic people: Openness predicts sensitivity to novelty in the experiences of interest and pleasure. *Frontiers in psychology*, 6:1877, 2015.
- [60] Mengjuan Fei, Wei Jiang, and Weijie Mao. Creating memorable video summaries that satisfy the user’s intention for taking the videos. *Neurocomputing*, 275:1911–1920, 2018.
- [61] Catrin Finkenauer, Rutger CME Engels, and Wim Meeus. Keeping secrets from parents: Advantages and disadvantages of secrecy in adolescence. *Journal of Youth and Adolescence*, 31(2):123–136, 2002.

- [62] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.
- [63] Barbara L. Fredrickson. *The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions.*, volume 56. American Psychological Association, 2001.
- [64] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.
- [65] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [66] Johan Galtung. Cultural violence. *Journal of peace research*, 27(3):291–305, 1990.
- [67] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41, 2016.
- [68] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In *Hellenic conference on artificial intelligence*, pages 91–100. Springer, 2010.
- [69] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [70] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, 50(2):1–36, 2017.
- [71] Yu Gong, Weiqiang Wang, Shuqiang Jiang, Qingming Huang, and Wen Gao. Detecting violent scenes in movies by auditory and visual cues. In *Pacific-Rim Conference on Multimedia*, pages 317–326. Springer, 2008.
- [72] Shinichi Goto and Terumasa Aoki. TUDCL at mediaeval 2013 violent scenes detection: Training with multi-modal features by MKL. In *Proceedings of the MediaEval 2013 Workshop*, 2013.
- [73] Helmut Grabner, Fabian Nater, Michel Druet, and Luc Van Gool. Visual interestingness in image sequences. In *ACM international conference on Multimedia*, pages 1017–1026. ACM, 2013.
- [74] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1633–1640, 2013.

- [75] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.
- [76] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3090–3098, 2015.
- [77] Michael Gygli and Mohammad Soleymani. Analyzing and predicting gif interestingness. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 122–126, 2016.
- [78] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, et al. Can we measure beauty? computational evaluation of coral reef aesthetics. *PeerJ*, 3:e1390, 2015.
- [79] Raisa Halonen, Stina Westman, and Pirkko Oittinen. Naturalness and interestingness of test images for visual quality evaluation. In *IS&T/SPIE Electronic Imaging*, pages 78670Z–78670Z. International Society for Optics and Photonics, 2011.
- [80] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. Learning computational models of video memorability from fmri brain imaging. *IEEE transactions on cybernetics*, 45(8):1692–1703, 2014.
- [81] Alex Hanson, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry Davis. Bidirectional convolutional lstm for the detection of violence in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [82] Judith M Harackiewicz, Jessi L Smith, and Stacy J Priniski. Interest matters: The importance of promoting interest in education. *Policy insights from the behavioral and brain sciences*, 3(2):220–227, 2016.
- [83] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE, 2012.
- [84] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [85] Alejandro Hernández-García, Fernando Fernández-Martínez, and Fernando Díaz-de María. Comparing visual descriptors and automatic rating strategies for video aesthetics prediction. *Signal Processing: Image Communication*, 47:280–288, 2016.

- [86] Eckhard H Hess and James M Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350, 1960.
- [87] Suzanne Hidi and Valerie Anderson. Situational interest and its impact on reading and expository writing. *The role of interest in learning and development*, 11:213–214, 1992.
- [88] Suzanne Hidi and William Baird. Interestingness—a neglected variable in discourse processing. *Cognitive science*, 10(2):179–194, 1986.
- [89] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [90] Liang-Chi Hsieh, Winston H Hsu, and Hao-Chuan Wang. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4309–4313. IEEE, 2014.
- [91] L Rowell Huesmann. The impact of electronic media violence: Scientific theory and research. *Journal of Adolescent health*, 41(6):S6–S13, 2007.
- [92] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in neural information processing systems*, pages 2429–2437, 2011.
- [93] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.
- [94] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011.
- [95] C. E. Izard and B. P. Ackerman. Emotion-cognition relationships and human development. *Lewis, Michael and Haviland-Jones, Jeannette M, Handbook of emotions*, pages 253–264, 2010.
- [96] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and predicting interestingness of videos. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.
- [97] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 73–79. ACM, 2014.
- [98] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168, 2015.



- [99] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Predicting subjectivity in image aesthetics assessment. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.
- [100] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [101] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE, 2006.
- [102] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.
- [103] Aditya Khosla, Jianxiang Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *Advances in neural information processing systems*, pages 296–304, 2012.
- [104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [105] Bert Krages. *Photography: the art of composition*. Skyhorse Publishing, Inc., 2012.
- [106] Bartosz Krawczyk, Leandro L Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.
- [107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [108] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.
- [109] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*, 1999.
- [110] Jieun Lee and Ilyoo B Hong. Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity. *International Journal of Information Management*, 36(3):360–373, 2016.

- [111] Roberto Leyva, Faiyaz Doctor, Alba G. Seco de Herrera, and Sohail Sahab. Multimodal deep features fusion for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [112] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing*, 3(2):236–252, 2009.
- [113] Xirong Li, Yujia Huo, Qin Jin, and Jieping Xu. Detecting violence in video using subclasses. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 586–590, 2016.
- [114] Cynthia CS Liem. Tud-mmcc at mediaeval 2016: Predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [115] Feng Liu, Yuzhen Niu, and Michael Gleicher. Using web photos for measuring video frame interestingness. In *Twenty-First International Joint Conference on Artificial Intelligence*, pages 2058–2063, 2009.
- [116] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92, 2010.
- [117] Gwenaëlle Marquant, Claire-Hélène Demarty, Christel Chamaret, Joël Sirot, and Louis Chevallier. Interestingness prediction & its application to immersive content. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2018.
- [118] Albert Mehrabian. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. 1980.
- [119] Shasha Mo, Jianwei Niu, Yiming Su, and Sajal K Das. A novel feature set for video emotion recognition. *Neurocomputing*, 291:11–20, 2018.
- [120] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.
- [121] Enrique Bermejo Nuevas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*, pages 332–339. Springer, 2011.
- [122] Pardis Noorzad and Bob L Sturm. Regression with sparse approximations of data. In *European Signal Processing Conference (EUSIPCO)*, pages 674–678. IEEE, 2012.

- [123] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [124] Balaji Padmanabhan and Alexander Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999.
- [125] Jayneel Parekh, Harshvardhan Tibrewal, and Sanjeel Parekh. Deep pairwise classification and ranking for predicting media interestingness. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 428–433, 2018.
- [126] Obioma Pelka, Christophe M Friedrich, A García Seco de Herrera, and Henning Müller. Overview of the imageclefmed 2019 concept detection task. *CLEF working notes, CEUR*, 2019.
- [127] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros. Technicolor and INRIA/IRISA at mediaeval 2011: learning temporal modality integration with bayesian networks. In *Proceedings of the MediaEval 2011 Workshop*, 2011.
- [128] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros. Technicolor-inria team at the mediaeval 2013 violent scenes detection task. In *Proceedings of the MediaEval 2013 Workshop*, 2013.
- [129] Cédric Penet, Claire-Hélène Demarty, Mohammad Soleymani, Guillaume Gravier, and Patrick Gros. Technicolor/inria/imperial college london at the mediaeval 2012 violent scene detection task. In *Working Notes Proceedings of the MediaEval 2012 Workshop.*, 2012.
- [130] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadvnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868. IEEE, 2015.
- [131] Reza Aditya Permadi, Septian Gilang Permana Putra, Cynthia Helmiriawan, and Cynthia CS Liem. Dut-mmsr at mediaeval 2017: Predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2017 Workshop.*, 2017.
- [132] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.
- [133] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and social psychology review*, 8(4):364–382, 2004.

- [134] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphaël Troncy, and Héctor Laria Mantecón. Combining textual and visual modeling for predicting media memorability. In *Working Notes Proceedings of the MediaEval 2019.*, 2019.
- [135] Miriam Redi and Bernard Merialdo. Where is the interestingness? retrieving appealing video scenes by learning flickr-based graded judgments. In *International Conference on Multimedia Retrieval*, pages 1363–1364. ACM, 2012.
- [136] Ronald A Rensink, J Kevin O’Regan, and James J Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, 8(5):368–373, 1997.
- [137] Lior Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12):4046–4072, 2009.
- [138] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 18–33. Springer Berlin Heidelberg, 2011.
- [139] Michael S Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*, 2019.
- [140] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.
- [141] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. Insiders and outsiders: Comparing urban impressions between population groups. In *ACM on International Conference on Multimedia Retrieval*, pages 65–71. ACM, 2017.
- [142] Samuel Felipe dos Santos and Jurandy Santos. Gibis at mediaeval 2019: Predicting media memorability task. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [143] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. Who’s afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM international conference on Multimedia*, pages 311–320. ACM, 2015.
- [144] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *AAAI Conference on Web and Social Media*, 2015.

- [145] Jan Schlüter, Bogdan Ionescu, Ionuț Mironică, and Markus Schedl. ARF @ mediaeval 2012: An uninformed approach to violence detection in hollywood movies. In *Proceedings of the MediaEval 2012 Workshop*, 2012.
- [146] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2009.
- [147] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. Umons at mediaeval 2015 affective impact of movies task including violent scenes detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*, 2015.
- [148] Amanda JC Sharkey. Types of multinet system. In *International Workshop on Multiple Classifier Systems*, pages 108–117. Springer, 2002.
- [149] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739, 2017.
- [150] Yuesong Shen, Claire-Hélène Demarty, and Ngoc Q. K. Duong. Technicolor@mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [151] Roger N Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 6(1):156–163, 1967.
- [152] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable? a deep style transfer approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 322–329, 2017.
- [153] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89–102, 2005.
- [154] Paul J Silvia. *Exploring the psychology of interest*. Oxford University Press, 2006.
- [155] Paul J Silvia. Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3(1):48, 2009.
- [156] Paul J Silvia and John B Warburton. Positive and negative affect: Bridging states and traits. *Comprehensive handbook of personality and psychopathology*, 1:268–284, 2006.



- [157] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [158] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. The mediaeval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*, 2015.
- [159] Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, Claire-Hélène Demarty, et al. The mediaeval 2014 affect task: Violent scenes detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*, 2014.
- [160] Mohammad Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 919–922, 2015.
- [161] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [162] Liviu-Daniel Ştefan, Mihai Gabriel Constantin, and Bogdan Ionescu. System fusion with deep ensembles. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 256–260, 2020.
- [163] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020.
- [164] Swathikiran Sudhakaran and Oswald Lanz. 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)*, pages 1–6. IEEE, 2017.
- [165] Jennifer J Sun, Ting Liu, and Gautam Prasad. Gla in mediaeval 2018 emotional impact of movies task. *arXiv preprint arXiv:1911.12361*, 2019.
- [166] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [167] Masaki Takahashi and Masanori Sano. Nhk where is beauty? grand challenge. In *ACM Multimedia challenge*, 2013.
- [168] Chun Chet Tan and Chong-Wah Ngo. The vireo team at mediaeval 2013: Violent scenes detection by mid-level concepts learnt from youtube. In *Proceedings of the MediaEval 2013 Workshop*, 2013.

- [169] Silvan S Tomkins. Affect, imagery, consciousness: Vol. i. the positive affects. 1962.
- [170] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European conference on computer vision*, pages 776–789. Springer, 2010.
- [171] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [172] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. Predicting media memorability using deep features with attention and recurrent network. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [173] Samuel A Turner Jr and Paul J Silvia. Must interesting things be pleasant? a test of competing appraisal structures. *Emotion*, 6(4):670, 2006.
- [174] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of experimental psychology: General*, 123(4):394–409, 1994.
- [175] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009.
- [176] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Eth-cvl@ mediaeval 2016: Textual-visual embeddings and video2gif for video interestingness. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016.
- [177] Alexander Viola and Sejong Yoon. A hybrid approach for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [178] Lijie Wang, Xueting Wang, and Toshihiko Yamasaki. Image aesthetics prediction using multiple patches preserving the original aspect ratio of contents. *arXiv preprint arXiv:2007.02268*, 2020.
- [179] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [180] Shuai Wang, Shizhe Chen, Jinming Zhao, and Qin Jin. 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*, pages 55–61, 2018.

- [181] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. Ruc at mediaeval 2019: Video memorability prediction based on visual textual and concept related features. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.
- [182] SHI Xingjian, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [183] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270, 2018.
- [184] Ying Xu, Yi Wang, Huaixuan Zhang, and Yong Jiang. Spatial attentive image aesthetic assessment. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
- [185] Wojtek Zajdel, Johannes D Krijnders, Tjeerd Andringa, and Darius M Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *2007 IEEE conference on advanced video and signal based surveillance*, pages 200–205. IEEE, 2007.
- [186] Nick Zangwill. Aesthetic judgment. In E. N. Zalta, editor, *The Stanford encyclopedia of philosophy*. Fall 2007 ed. edition, 2003.
- [187] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3320. IEEE, 2011.
- [188] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM international conference on Multimedia*, pages 47–56. ACM, 2014.

# Automatic analysis of the visual impact of multimedia data

## ORIGINALITY REPORT

8%

SIMILARITY INDEX

5%

INTERNET SOURCES

5%

PUBLICATIONS

1%

STUDENT PAPERS

## PRIMARY SOURCES

- 1** Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, Bogdan Ionescu. "Computational Understanding of Visual Interestingness Beyond Semantics", ACM Computing Surveys, 2019  
Publication 1%
- 2** [ceur-ws.org](http://ceur-ws.org)  
Internet Source 1%
- 3** [hal.archives-ouvertes.fr](http://hal.archives-ouvertes.fr)  
Internet Source 1%
- 4** Mihai Gabriel Constantin, Bogdan Ionescu. "Content description for Predicting image Interestingness", 2017 International Symposium on Signals, Circuits and Systems (ISSCS), 2017  
Publication 1%
- 5** Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Helene Demarty et al. "Affect in Multimedia: Benchmarking Violent Scenes Detection", IEEE Transactions on Affective Computing, 2020  
Publication <1%

6

Yashar Deldjoo, Markus Schedl. "Retrieving Relevant and Diverse Movie Clips Using the MFVCD-7K Multifaceted Video Clip Dataset", 2019 International Conference on Content-Based Multimedia Indexing (CBMI), 2019

Publication

<1%

7

[export.arxiv.org](https://export.arxiv.org)

Internet Source

<1%

8

Submitted to University of Southern California

Student Paper

<1%

9

[researchr.org](https://researchr.org)

Internet Source

<1%

10

[www.eurecom.fr](https://www.eurecom.fr)

Internet Source

<1%

11

[hal.inria.fr](https://hal.inria.fr)

Internet Source

<1%

12

Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, Mats Sjöberg. "Datasets column", ACM SIGMultimedia Records, 2019

Publication

<1%

13

Submitted to Technical University of Cluj-Napoca

Student Paper

<1%

14

[www.mdpi.com](https://www.mdpi.com)



Internet Source

<1%

15

[livrepository.liverpool.ac.uk](http://livrepository.liverpool.ac.uk)

Internet Source

<1%

16

[repositorio.unicamp.br](http://repositorio.unicamp.br)

Internet Source

<1%

17

[summit.sfu.ca](http://summit.sfu.ca)

Internet Source

<1%

18

[en.d2l.ai](http://en.d2l.ai)

Internet Source

<1%

19

"Visual Content Indexing and Retrieval with Psycho-Visual Models", Springer Science and Business Media LLC, 2017

Publication

<1%

20

Submitted to University of Pretoria

Student Paper

<1%

21

Olfa Ben-Ahmed, Benoit Huet. "Deep Multimodal Features for Movie Genre and Interestingness Prediction", 2018 International Conference on Content-Based Multimedia Indexing (CBMI), 2018

Publication

<1%

22

[dblp.dagstuhl.de](http://dblp.dagstuhl.de)

Internet Source

<1%

23

Internet Source

&lt;1%

24

[www.jku.at](http://www.jku.at)

Internet Source

&lt;1%

25

[arxiv.org](http://arxiv.org)

Internet Source

&lt;1%

26

Romain Cohendet, Claire-Helene Demarty, Ngoc Duong, Martin Engilberge. "VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019

Publication

&lt;1%

27

"Frontier Computing", Springer Science and Business Media LLC, 2020

Publication

&lt;1%

28

[d-nb.info](http://d-nb.info)

Internet Source

&lt;1%

29

[orca.cf.ac.uk](http://orca.cf.ac.uk)

Internet Source

&lt;1%

30

Submitted to Universiti Teknologi Malaysia

Student Paper

&lt;1%

31

[escholarship.org](http://escholarship.org)

Internet Source

&lt;1%

"Advances in Multimedia Information Processing

32

– PCM 2018", Springer Science and Business  
Media LLC, 2018

Publication

<1%

33

[digitalcommons.njit.edu](https://digitalcommons.njit.edu)

Internet Source

<1%

34

Submitted to National University of Singapore

Student Paper

<1%

35

"Advances in Information Retrieval", Springer  
Science and Business Media LLC, 2020

Publication

<1%

36

"ECAI 2020", IOS Press, 2020

Publication

<1%

37

Sicheng Zhao, Shangfei Wang, Mohammad  
Soleymani, Dhiraj Joshi, Qiang Ji. "Affective  
Computing for Large-scale Heterogeneous  
Multimedia Data", ACM Transactions on  
Multimedia Computing, Communications, and  
Applications, 2019

Publication

<1%

38

Submitted to October University for Modern  
Sciences and Arts (MSA)

Student Paper

<1%

39

[sejongyoon.net](http://sejongyoon.net)

Internet Source

<1%

40

[yanweifu.github.io](https://yanweifu.github.io)

Internet Source

<1%

---

41

[dspace.vsb.cz](https://dspace.vsb.cz)

Internet Source

<1%

---

42

Advances in Computer Vision and Pattern Recognition, 2015.

Publication

<1%

---

43

[www.citeulike.org](http://www.citeulike.org)

Internet Source

<1%

---

44

[www.researchgate.net](http://www.researchgate.net)

Internet Source

<1%

---

45

"MultiMedia Modeling", Springer Science and Business Media LLC, 2020

Publication

<1%

---

46

[www.iieta.org](http://www.iieta.org)

Internet Source

<1%

---

47

[eprint.iacr.org](http://eprint.iacr.org)

Internet Source

<1%

---

48

[eprints-phd.biblio.unitn.it](http://eprints-phd.biblio.unitn.it)

Internet Source

<1%

---

49

[www.preference-learning.org](http://www.preference-learning.org)

Internet Source

<1%

---

50

Lecture Notes in Computer Science, 2012.

Publication

<1%

---

51

[dl.acm.org](http://dl.acm.org)

Internet Source

<1%

---

52	<a href="http://upcommons.upc.edu">upcommons.upc.edu</a> Internet Source	<1%
53	<a href="http://repository.ntu.edu.sg">repository.ntu.edu.sg</a> Internet Source	<1%
54	<a href="http://ruidera.uclm.es">ruidera.uclm.es</a> Internet Source	<1%
55	<a href="http://eprints.nottingham.ac.uk">eprints.nottingham.ac.uk</a> Internet Source	<1%
56	<a href="http://bmcmeginformdecismak.biomedcentral.com">bmcmeginformdecismak.biomedcentral.com</a> Internet Source	<1%
57	<a href="http://dspace.cuni.cz">dspace.cuni.cz</a> Internet Source	<1%
58	<a href="http://tel.archives-ouvertes.fr">tel.archives-ouvertes.fr</a> Internet Source	<1%
59	Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, Claire-Hélène Demarty. "Annotating, Understanding, and Predicting Long-term Video Memorability", Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval - ICMR '18, 2018 Publication	<1%
60	<a href="http://manualzz.com">manualzz.com</a> Internet Source	<1%
61	Yang Liu, Zhonglei Gu, Yiu-ming Cheung, Kien	<1%



A. Hua. "Multi-view Manifold Learning for Media Interestingness Prediction", Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval - ICMR '17, 2017

Publication

62

[link.springer.com](http://link.springer.com)

Internet Source

<1%

63

[www.tandfonline.com](http://www.tandfonline.com)

Internet Source

<1%

64

[repository.dl.itc.u-tokyo.ac.jp](http://repository.dl.itc.u-tokyo.ac.jp)

Internet Source

<1%

65

[dahualin.org](http://dahualin.org)

Internet Source

<1%

66

[www.ittrend.co.kr](http://www.ittrend.co.kr)

Internet Source

<1%

67

Eloise Berson, Ngoc Q.K. Duong, Claire-Helene Demarty. "Collecting, Analyzing and Predicting Socially-Driven Image Interestingness", 2019 27th European Signal Processing Conference (EUSIPCO), 2019

Publication

<1%

68

Jurandy Almeida, Lucas P. Valem, Daniel C. G. Pedronette. "Chapter 1 A Rank Aggregation Framework for Video Interestingness Prediction", Springer Science and Business Media LLC, 2017

<1%

69

Multimodal Location Estimation of Videos and Images, 2015.

Publication

<1%

---

70

"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020

Publication

<1%

---

Exclude quotes      On

Exclude matches      < 5 words

Exclude bibliography      On

## Document Viewer

## Turnitin Originality Report

Processed on: 23-Oct-2020 15:26 EEST

ID: 1424159736

Word Count: 28710

Submitted: 1

Automatic analysis of the visual impact of  
mu... By Mihai Gabriel Constantin

Similarity Index


8%

## Similarity by Source


Internet Sources:	5%
Publications:	5%
Student Papers:	1%

[include quoted](#)
[include bibliography](#)
[excluding matches < 5 words](#)
mode:  
[Change mode](#)
[print](#)
[refresh](#)
[download](#)

1% match (publications)

[Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, Bogdan Ionescu. "Computational Understanding of Visual Interestingness Beyond Semantics", ACM Computing Surveys, 2019](#) 

1% match (publications)

[Mihai Gabriel Constantin, Bogdan Ionescu. "Content description for Predicting image Interestingness", 2017 International Symposium on Signals, Circuits and Systems \(ISSCS\), 2017](#) 

&lt;1% match ()

<https://hal.archives-ouvertes.fr/hal-02368920> 

&lt;1% match (publications)

[Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Helene Demarty et al. "Affect in Multimedia: Benchmarking Violent Scenes Detection", IEEE Transactions on Affective Computing, 2020](#) 

&lt;1% match (Internet from 27-Nov-2016)

<http://ceur-ws.org> 

<1% match (Internet from 07-Nov-2017) <a href="http://ceur-ws.org">http://ceur-ws.org</a>	✕
<1% match (Internet from 10-Sep-2017) <a href="https://hal.archives-ouvertes.fr/hal-01497425/document">https://hal.archives-ouvertes.fr/hal-01497425/document</a>	✕
<1% match (publications) <a href="#">Yashar Deldjoo, Markus Schedl. "Retrieving Relevant and Diverse Movie Clips Using the MFVCD-7K Multifaceted Video Clip Dataset", 2019 International Conference on Content-Based Multimedia Indexing (CBMI), 2019</a>	✕
<1% match (student papers from 22-Sep-2013) <a href="#">Submitted to University of Southern California on 2013-09-22</a>	✕
<1% match (Internet from 22-May-2020) <a href="http://www.eurecom.fr">http://www.eurecom.fr</a>	✕
<1% match (publications) <a href="#">Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, Mats Sjöberg. "Datasets column", ACM SIGMultimedia Records, 2019</a>	✕
<1% match (student papers from 01-Mar-2016) <a href="#">Submitted to Technical University of Cluj-Napoca on 2016-03-01</a>	✕
<1% match (Internet from 08-Jul-2020) <a href="https://researchr.org/publication/CarataCGCG18">https://researchr.org/publication/CarataCGCG18</a>	✕
<1% match (Internet from 05-Oct-2020) <a href="https://www.mdpi.com/2072-4292/9/2/133/htm">https://www.mdpi.com/2072-4292/9/2/133/htm</a>	✕
<1% match (Internet from 23-Apr-2020) <a href="http://repositorio.unicamp.br">http://repositorio.unicamp.br</a>	✕
<1% match (Internet from 19-Mar-2020) <a href="http://summit.sfu.ca">http://summit.sfu.ca</a>	✕
<1% match (Internet from 01-Oct-2019) <a href="https://en.d2l.ai/d2l-en.pdf">https://en.d2l.ai/d2l-en.pdf</a>	✕
<1% match (publications) <a href="#">"Visual Content Indexing and Retrieval with Psycho-Visual Models", Springer Science and Business Media LLC, 2017</a>	✕

<1% match (publications) <a href="#">Olfa Ben-Ahmed, Benoit Huet. "Deep Multimodal Features for Movie Genre and Interestingness Prediction", 2018 International Conference on Content-Based Multimedia Indexing (CBMI), 2018</a>	✕
<1% match (student papers from 12-Nov-2019) <a href="#">Submitted to University of Pretoria on 2019-11-12</a>	✕
<1% match (Internet from 07-Jul-2020) <a href="https://dblp.dagstuhl.de/search/publ/bibtex/?q=stream%3Astreams%2Fconf%2Fmediaeval%3A">https://dblp.dagstuhl.de/search/publ/bibtex/?q=stream%3Astreams%2Fconf%2Fmediaeval%3A</a>	✕
<1% match (Internet from 10-Aug-2020) <a href="https://livrepository.liverpool.ac.uk/3022573/1/201004595_June2018.pdf">https://livrepository.liverpool.ac.uk/3022573/1/201004595_June2018.pdf</a>	✕
<1% match (Internet from 30-Jul-2020) <a href="https://hal.inria.fr/hal-02570804v2/document">https://hal.inria.fr/hal-02570804v2/document</a>	✕
<1% match (Internet from 21-May-2020) <a href="https://www.jku.at/fileadmin/gruppen/173/Research/deldjoo_mmsys_2018.pdf">https://www.jku.at/fileadmin/gruppen/173/Research/deldjoo_mmsys_2018.pdf</a>	✕
<1% match (Internet from 24-Oct-2018) <a href="https://d-nb.info/115627690X/34">https://d-nb.info/115627690X/34</a>	✕
<1% match (Internet from 17-Aug-2020) <a href="https://arxiv.org/pdf/1812.01973.pdf">https://arxiv.org/pdf/1812.01973.pdf</a>	✕
<1% match () <a href="http://orca.cf.ac.uk">http://orca.cf.ac.uk</a>	✕
<1% match (Internet from 05-Sep-2019) <a href="https://escholarship.org/content/qt4bj8540h/qt4bj8540h.pdf?t=pwztia">https://escholarship.org/content/qt4bj8540h/qt4bj8540h.pdf?t=pwztia</a>	✕
<1% match (publications) <a href="#">"Frontier Computing", Springer Science and Business Media LLC, 2020</a>	✕
<1% match (publications) <a href="#">Romain Cohendet, Claire-Helene Demarty, Ngoc Duong, Martin Engilberge. "VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019</a>	✕
<1% match (student papers from 26-Jul-2016) <a href="#">Submitted to Universiti Teknologi Malaysia on 2016-07-26</a>	



<1% match (Internet from 07-Nov-2017) <a href="http://ceur-ws.org">http://ceur-ws.org</a>	✕
<1% match (Internet from 24-Jul-2020) <a href="https://digitalcommons.njit.edu/cgi/viewcontent.cgi?amp=&amp;article=1019&amp;context=dissertations">https://digitalcommons.njit.edu/cgi/viewcontent.cgi?amp=&amp;article=1019&amp;context=dissertations</a>	✕
<1% match (publications) <a href="#">"Advances in Multimedia Information Processing – PCM 2018", Springer Science and Business Media LLC, 2018</a>	✕
<1% match () <a href="https://hal.archives-ouvertes.fr/hal-02285826/file/PID5990011_cameraready.pdf">https://hal.archives-ouvertes.fr/hal-02285826/file/PID5990011_cameraready.pdf</a>	✕
<1% match (publications) <a href="#">"Advances in Information Retrieval", Springer Science and Business Media LLC, 2020</a>	✕
<1% match (publications) <a href="#">"ECAI 2020", IOS Press, 2020</a>	✕
<1% match (student papers from 26-Aug-2014) <a href="#">Submitted to National University of Singapore on 2014-08-26</a>	✕
<1% match (Internet from 09-Oct-2020) <a href="https://researchr.org/publication/ConstantinKDDVI19">https://researchr.org/publication/ConstantinKDDVI19</a>	✕
<1% match (publications) <a href="#">Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, Qiang Ji. "Affective Computing for Large-scale Heterogeneous Multimedia Data", ACM Transactions on Multimedia Computing, Communications, and Applications, 2019</a>	✕
<1% match (student papers from 07-Apr-2019) <a href="#">Submitted to October University for Modern Sciences and Arts (MSA) on 2019-04-07</a>	✕
<1% match (Internet from 14-Jul-2020) <a href="http://sejongyoon.net">http://sejongyoon.net</a>	✕
<1% match (Internet from 17-Apr-2019) <a href="http://yanweifu.github.io">http://yanweifu.github.io</a>	✕
<1% match (Internet from 08-Apr-2019) <a href="http://dspace.vsb.cz">http://dspace.vsb.cz</a>	✕

<1% match (Internet from 28-Sep-2011) <a href="http://www.citeulike.org">http://www.citeulike.org</a>	✕
<1% match (publications) <a href="#">Advances in Computer Vision and Pattern Recognition, 2015.</a>	✕
<1% match (Internet from 27-Aug-2020) <a href="https://www.researchgate.net/publication/259037619_THE_MILLION_MUSICAL_TWEETS_DATASET_WHAT_CAN_WE_LEARN_FROM_MICROBLOGS">https://www.researchgate.net/publication/259037619_THE_MILLION_MUSICAL_TWEETS_DATASET_WHAT_CAN_WE_LEARN_FROM_MICROBLOGS</a>	✕
<1% match (Internet from 18-Oct-2019) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 22-Jul-2020) <a href="https://hal.inria.fr/hal-02366687/file/srijan_iccv19.pdf">https://hal.inria.fr/hal-02366687/file/srijan_iccv19.pdf</a>	✕
<1% match (Internet from 16-May-2018) <a href="https://www.diva-portal.org/smash/get/diva2:561201/FULLTEXT01.pdf">https://www.diva-portal.org/smash/get/diva2:561201/FULLTEXT01.pdf</a>	✕
<1% match (publications) <a href="#">"MultiMedia Modeling", Springer Science and Business Media LLC, 2020</a>	✕
<1% match () <a href="http://livrepository.liverpool.ac.uk">http://livrepository.liverpool.ac.uk</a>	✕
<1% match (Internet from 20-Oct-2020) <a href="http://www.iieta.org">http://www.iieta.org</a>	✕
<1% match (Internet from 19-Jan-2020) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 07-Jun-2020) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 23-Apr-2020) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 06-Nov-2018) <a href="https://eprint.iacr.org/2018/053.pdf">https://eprint.iacr.org/2018/053.pdf</a>	✕
<1% match (Internet from 06-Sep-2017)	

<a href="http://eprints-phd.biblio.unitn.it">http://eprints-phd.biblio.unitn.it</a>	✕
<1% match (Internet from 01-Sep-2014) <a href="http://www.preference-learning.org">http://www.preference-learning.org</a>	✕
<1% match (publications) <a href="#">Lecture Notes in Computer Science, 2012.</a>	✕
<1% match (Internet from 24-May-2014) <a href="http://www.diva-portal.org">http://www.diva-portal.org</a>	✕
<1% match (Internet from 06-Oct-2020) <a href="https://ruidera.uclm.es/xmlui/bitstream/handle/10578/12481/TESIS%20Serrano%20Gracia.pdf?isAllowed=y&amp;sequence=1">https://ruidera.uclm.es/xmlui/bitstream/handle/10578/12481/TESIS%20Serrano%20Gracia.pdf?isAllowed=y&amp;sequence=1</a>	✕
<1% match (Internet from 11-Feb-2019) <a href="http://eprints.nottingham.ac.uk">http://eprints.nottingham.ac.uk</a>	✕
<1% match (Internet from 06-Jan-2020) <a href="https://upcommons.upc.edu/bitstream/handle/2117/107669/predicting-media-interestingness.pdf?isAllowed=y&amp;sequence=1">https://upcommons.upc.edu/bitstream/handle/2117/107669/predicting-media-interestingness.pdf?isAllowed=y&amp;sequence=1</a>	✕
<1% match (Internet from 16-Mar-2019) <a href="http://dahualin.org">http://dahualin.org</a>	✕
<1% match (Internet from 27-May-2019) <a href="https://hal.inria.fr/hal-02140558/document">https://hal.inria.fr/hal-02140558/document</a>	✕
<1% match (Internet from 02-Jul-2019) <a href="https://dspace.cuni.cz/bitstream/handle/20.500.11956/107024/120329717.pdf?isAllowed=y&amp;sequence=1">https://dspace.cuni.cz/bitstream/handle/20.500.11956/107024/120329717.pdf?isAllowed=y&amp;sequence=1</a>	✕
<1% match (Internet from 31-Aug-2017) <a href="https://repository.ntu.edu.sg/bitstream/handle/10356/69073/Thesis_final.pdf?isAllowed=n&amp;sequence=1">https://repository.ntu.edu.sg/bitstream/handle/10356/69073/Thesis_final.pdf?isAllowed=n&amp;sequence=1</a>	✕
<1% match (Internet from 07-Aug-2019) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 19-Feb-2017) <a href="http://dl.acm.org">http://dl.acm.org</a>	✕
<1% match ( )	

<a href="https://repository.dl.itc.u-tokyo.ac.jp/?action=repository_action_common_download&amp;item_id=50267&amp;item_no=1&amp;attribute_id=14&amp;file_no=2">https://repository.dl.itc.u-tokyo.ac.jp/?action=repository_action_common_download&amp;item_id=50267&amp;item_no=1&amp;attribute_id=14&amp;file_no=2</a>	✕
<1% match (Internet from 10-Apr-2018) <a href="https://hal.archives-ouvertes.fr/tel-01449813/file/Thesis_Prasad_Samarakoon.pdf">https://hal.archives-ouvertes.fr/tel-01449813/file/Thesis_Prasad_Samarakoon.pdf</a>	✕
<1% match (Internet from 06-Feb-2020) <a href="https://www.tandfonline.com/doi/full/10.1080/02699931.2015.1071241">https://www.tandfonline.com/doi/full/10.1080/02699931.2015.1071241</a>	✕
<1% match (Internet from 14-Aug-2020) <a href="https://tel.archives-ouvertes.fr/tel-02492463/document">https://tel.archives-ouvertes.fr/tel-02492463/document</a>	✕
<1% match (Internet from 22-May-2020) <a href="http://export.arxiv.org">http://export.arxiv.org</a>	✕
<1% match (Internet from 19-Mar-2020) <a href="https://manualzz.com/doc/32890211/multimodal-surveillance-behavior-analysis-for-recognizing">https://manualzz.com/doc/32890211/multimodal-surveillance-behavior-analysis-for-recognizing</a>	✕
<1% match (Internet from 01-May-2020) <a href="https://bmcmidinformatik.biomedcentral.com/articles/10.1186/s12911-019-0761-8">https://bmcmidinformatik.biomedcentral.com/articles/10.1186/s12911-019-0761-8</a>	✕
<1% match (Internet from 02-Dec-2019) <a href="https://link.springer.com/content/pdf/10.1007%2F978-3-319-55753-3.pdf">https://link.springer.com/content/pdf/10.1007%2F978-3-319-55753-3.pdf</a>	✕
<1% match (Internet from 20-Oct-2013) <a href="http://www.ittrend.co.kr">http://www.ittrend.co.kr</a>	✕
<1% match (publications) <a href="#">Yang Liu, Zhonglei Gu, Yiu-ming Cheung, Kien A. Hua. "Multi-view Manifold Learning for Media Interestingness Prediction", Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval - ICMR '17, 2017</a>	✕
<1% match (publications) <a href="#">Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, Claire-Hélène Demarty. "Annotating, Understanding, and Predicting Long-term Video Memorability", Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval - ICMR '18, 2018</a>	✕
<1% match (publications) <a href="#">Eloise Berson, Ngoc Q.K. Duong, Claire-Helene Demarty. "Collecting, Analyzing and Predicting Socially-Driven Image Interestingness", 2019 27th European Signal Processing Conference (EUSIPCO), 2019</a>	✕
<1% match (publications) <a href="#">Jurandy Almeida, Lucas P. Valem, Daniel C. G. Pedronette. "Chapter 1 A Rank Aggregation Framework for Video</a>	

[Interestingness Prediction", Springer Science and Business Media LLC, 2017](#)

<1% match (publications)

[Multimodal Location Estimation of Videos and Images, 2015.](#)

<1% match (publications)

["Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020](#)

“POLITEHNICA” UNIVERSITY OF BUCHAREST ETTI-B DOCTORAL SCHOOL Decision No. 569 from 25.09.2020 Automatic analysis of the visual impact of multimedia data Analiza automată a impactului vizual al datelor multimedia by Mihai Gabriel Constantin [A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electronics and Telecommunications COMISIA DE DOCTORAT Președinte Prof. Dr. Ing. Gheroghe Brezeanu de la Universitatea Politehnica București Conducător de doctorat Prof. Dr. Ing. Bogdan Ionescu de la Universitatea Politehnica București Referent Prof. Dr. Ing. Martha Larson de la Radbound University Olanda Referent Dr. Ing. Claire-Hélène Demarty de la InterDigital, Franța Referent Prof. Dr. Ing. Mihai Ciuc de la Universitatea Politehnica București](#) Bucharest 2020 [Acknowledgements I would firstly like to thank my](#) thesis coordinator, Prof. Dr. Ing. Bogdan Ionescu, [for his patience and guidance during the time I spent working on this](#) thesis. [I would also like to thank](#) him for introducing me to this domain and to the scientific community that revolves around it. With his help, I discovered the challenging, ever- evolving, and fascinating world of academia and research. As my doctoral program reaches this stage, I can only hope for a long and fruitful future collaboration with him. I would also like to thank the MediaEval community and the person who drives and organizes this community, Prof. Dr. Ing. Martha Larson. The benchmarking tasks published in this community represent one of the pillars upon which this thesis is created. Also, a big thank you to my colleagues in the Multimedia Lab for their help, input, and collaboration on the research projects and paper we published together. My gratitude also goes to my friends, who encouraged me to follow a career in academia. Finally, [I would like to thank my parents for their continuous help, support, and encouragement in all my endeavors.](#)

iii Contents Acknowledgements . . . . .  
. . . . . [List of abbreviations](#) . . . . . [1 Introduction](#) [1.1 Domain of the thesis](#) . . . . . [1.2 Motivation of the thesis](#) . . . . .  
. [1.3 Content of the thesis](#) . . . . . 2 Theoretical aspects [2.1 Taxonomy and definitions](#) . . . . . [2.2 Human understanding of the subjective properties of multimedia data](#) [2.3 Datasets and user studies](#) . . . . . [2.4 Computational approaches](#) . . . . . [2.4.1 Interestingness](#) . . . . .  
[2.4.2 Aesthetic value](#) . . . . . [2.4.3 Memorability](#) . . . . .  
. . . . . [2.4.4 Violence](#) . . . . . [2.4.5 Affective value and emotions.](#) . . . . .  
. . . . . [2.5 Applications](#) . . . . . [2.6 Conclusions](#) . . . . .  
. . . . . v iii ii 1 2 4 5 6 10 13 17 17 19 20 21 22 23 25 3 Personal contributions [3.1 Datasets and evaluation](#) . . . . . [3.1.1 Interestingness prediction](#) . . . . . [3.1.2 Violence prediction](#) . . . . . [3.1.3](#)



Memorability prediction . . . . .	3.1.4 Content recommendation . . . . .
. . . . . 3.2 Predicting media interestingness . . . . .	3.2.1 Introduction . . . . .
. . . . . 3.2.2 SVM-based learning systems . . . . .	3.2.3 Aesthetic features and late fusion learning systems . . . . .
. . . . . 3.2.4 Conclusions . . . . .	
. 3.3 Predicting violent scenes . . . . .	3.3.1 Introduction . . . . .
. . . . . 3.3.2 Temporal deep learning systems . . . . .	3.3.3 Conclusions . . . . .
. . . . . 3.4 Predicting media memorability . . . . .	3.4.1 Introduction . . . . .
. . . . . 3.4.2 Action-based deep learning systems . . . . .	
. . . . . 3.4.3 Conclusions . . . . .	3.5 Late fusion with deep ensemble systems . . . . .
. . . . . 3.5.1 Introduction . . . . .	3.5.2 Motivation . . . . .
. . . . . 3.5.3 Previous work . . . . .	
. . . . . 3.5.4 Proposed approach . . . . .	3.5.5 Experimental setup . . . . .
. . . . . 3.5.6 Experimental results . . . . .	3.5.7 Conclusions . . . . .
. . . . . vi 27 27 27 35 39 41 43 43 43 48 54 57 57 57 60 61 61 61 65 67 67 67 68 68 75 77 81 4	4 General conclusions and perspectives 4.1 Contributions and publications . . . . .
. . . . . 4.2 Conclusions . . . . .	4.3 Future perspectives . . . . .
. . . . . Bibliography 83 83 90 91 93 vii	List of abbreviations 1D, 2D, 3D - one-, two-, three-dimensional BN - Batch normalization CNN - Convolutional neural network CSF - Cross-Space-Fusion layer DNN - Deep neural network HOG - Histogram of oriented gradients HSV - Hue-saturation-value IQR - Interquartile range KF - K-fold kNN - k-nearest neighbors algorithm LBP - Local binary patterns LOF - Local outlier factor LSTM - Long short-term memory MAP - Mean average precision <a href="#">MLP - Multi-layered perceptron</a> <a href="#">RBF - Radial basis function</a> <a href="#">SIFT - Scale-invariant feature transform</a> <a href="#">SVM - Support vector machine</a> VAD - Valence-arousal-dominance ii
Chapter 1 Intro duction 1.1 Domain of the thesis This thesis presents and analyzes several aspects and <a href="#">state of the art methods</a> that cover <a href="#">the automatic analysis of the</a> visual impact of multimedia data, with an accent on the study of a number essential concepts in this domain, such as interesting- ness, aesthetics, memorability, violence, and affective value and emotions. While more traditional computer vision tasks attempt to solve problems that have objective ground-truth values that all or most annotators would agree with, such as object detection or scene classification, recent developments in deep neural network pro- cessing, social media, hardware availability and cost, psychological studies, and big data availability allowed scientists to expand their research into domains that target more subjective concepts. In the latter case, ground truth may depend on a large number of human-centric factors, including, but not limited to, personal preferences, cultural background, cognitive abilities, and current psychological state. Predicting and understanding such concepts with the help of computer vision methods dramat- ically increases the utility and added value created by implementing such methods, allowing scientists and developers to predict how multimedia data affects viewers. 1 However, the development of such methods is not trivial. Researchers from many dif- ferent domains must be involved and must work together in order to create accurate predictors that can function in a mostly online, real-world environment that deals with large amounts of diverse visual data. Researchers in cognitive and humanities sciences, physiologists, specialists in human behavior, human data annotation, and computer vision algorithm developers must come together in order to define these concepts, create theories about how they influence perception and behavior, collect and annotate a large	

amount of data and create the computer vision methods that can predict the concepts. In this thesis, I present a literature survey on my main concepts of interest for my field of study, which predominantly revolves around interestingness, aesthetic value, memorability, violence and affective value and emotional content, and continue with presenting my main contributions, both to the collection of datasets and the creation of common evaluation benchmarks, to computer vision methods for the prediction of such concepts, as well as the creation of a novel deep learning-based late fusion system, that significantly increases the performance of its inducer systems. All these contributions are developed during my Ph.D. studies.

### 1.2 Motivation of the thesis

This thesis aims to contribute to the understanding of such subjective concepts, study, discover, and underline the current best practices and best-performing methods and models for certain tasks, and create computer vision methods that successfully predict the targeted concepts. Given the advent of social media, increasingly larger collections of images and videos are available for users, and it becomes increasingly difficult to navigate them. On the positive side, access to a larger amount of data can be beneficial for system development, as more training and testing samples are available, especially for deep neural networks, that are known for their high demand for annotated data. While, as previously shown, this extensive collection of concepts present varying degrees of subjectivity, and, therefore, inter- and even intra-rater reliability with regards to the annotated image and video samples in given datasets can significantly vary, the interest for computer vision methods that solve these problems and predict these concepts is growing, regardless of the difficulties created by the inherent concept subjectivity. From this perspective, there is a large demand for these methods, mostly driven by social media, media sharing, advertising, and media archiving platforms. These particular branches of industry would benefit from the creation of automatic predictors, recommender systems based on these concepts, automatic filters, and other functionalities that would be impossible to implement without the help of computer vision, machine learning, and artificial intelligence. Currently, some of these concepts are starting to be implemented in professional solutions and web services. Pioneers in this direction are represented by popular websites and social media platforms, like Flickr 1, who implement a social interestingness-based metric for creating suggestions with regards to new images and posts, or Google Photos 2, that can create short summaries of photos based on the appeal of photos uploaded by users in their personal collections. Support from the industry is also manifested by the support for particular tasks that aid both the research community and the industry, such as InterDigital's support for the study of multimedia interestingness, memorability, violence, and emotional content prediction 3. Thus, researchers that create tasks, datasets, and computational models are motivated to keep in mind and target realistic use case scenarios that can be implemented in such environments. [1https://www.flickr.com/](https://www.flickr.com/) [2https://www.google.com/photos/about/](https://www.google.com/photos/about/) [3https://www.interdigital.com/datasets/](https://www.interdigital.com/datasets/)

### 1.3 Content of the thesis

The rest of this thesis is divided into 3 Chapters. The first one presents the current state-of-the-art with regards to taxonomies, psychological studies, datasets, user studies, and computational approaches developed by researchers from different domains that handle the problem of defining and predicting the subjective properties of multimedia data. The second chapter presents personal contributions to this domain, with regards to the datasets and evaluation benchmarks I helped create, and to original computational methods and models for the prediction of some of these concepts, as well as a generalized deep learning-based collection of late fusion approaches that can accurately predict the given concepts, using a large selection of weaker input

inducers. The thesis ends with some general conclusions and perspectives for future works, as well as a summary of my papers and contribution to those papers. Chapter 2 Theoretical aspects In today's internet and big data landscape, users are constantly bombarded with large quantities of multimedia data, sometimes creating that data themselves via personal photo collections, social media posts, or personal vlogs. It is indeed difficult to keep track of all that information. Researchers have shown that this constant feed of information, both visual and otherwise, can significantly reduce the human attention span [138]. This environment creates the need for the development of systems that would help human users navigate this tremendous amount of data, whether we are talking about systems aimed at sorting data, based on how interesting, appealing, or memorable it is, systems aimed at creating filters capable of detecting violent or emotionally scarring data, or recommending other media samples that are more in tuned with personal user preferences. One of the hardest challenges these systems face is represented by the definitions of these concepts, considering that, unlike more tangible tasks such as detecting an object in an image, most of the times, it is hard for human subjects to agree on what is interesting, aesthetically pleasing, violent, and so on. The subjective nature of these proprieties does make their prediction and classification one of the more challenging tasks in computer vision today. A close collaboration between theorists in humanities, human behavior, and computer vision, is therefore necessary in order to create algorithms and market-ready systems. This chapter will present a literature review and analysis focused on concepts that will be used throughout the thesis, namely interestingness, aesthetics, memorability, violence, and affective value and emotions.

2.1 Taxonomy and definitions As previously mentioned, the first important step in analyzing these targeted concepts is creating a list of possible definitions for them, having as starting point psychological theories, applied human studies, and use-case scenarios/ An extensive set of subjective proprieties has been studied in the current literature. As we present in [33], some taxonomies can be built in order to understand and classify these concepts. Table 2.1 presents a list of subjective concepts studied by scientists, grouped according to a central common theme. For example, novelty, originality, unexpectedness, etc., tend to measure the novelty of media samples from different perspectives and therefore fall into the same central theme. Another possible approach regarding the creation of taxonomies is an analysis of concept correlation. In this case, having just one target concept as a starting point, a list of correlation with other concepts can be created based on research works in the current state-of-the-art literature. These correlations can be positive, negative, or un- defined (or mostly not explored). Such an example is presented in our work [33] and in Table 2.2, where a taxonomy based on correlation with interestingness is presented. The table represents a thorough analysis that considers papers from psychological, user study-based, or computational perspectives. Furthermore, an example of a sci- entific paper that studies the correlation is given for each concept. Even in such a thorough analysis, some controversies arise that show certain concepts **to be both positively and negatively correlated with interestingness**. Such issues may occur due

Table 2.1: Taxonomy. List of concepts that are covered in the current **state-of-the-art** literature **as** presented **in** Constantin **et al.** [33]. Concepts in this table are grouped according to a central theme. Theme **Close concepts** 1 **Interestingness** **Interestingness** 2 **Affective Value and Emotions** Dimensional Emotion Space (Valence/Pleasantness, Arousal, Dominance) and Categorical Emotion Space (Happiness, Boredom, etc.) 3 **Aesthetic Value** **Aesthetic Value and Cuteness** 4 **Memorability** Memorability 5 **Novelty** Novelty, Originality, Unusualness, Unexpectedness, Distinctiveness and Familiarity 6 **Complexity** Complexity and Simplicity 7 **Coping**

Potential Coping Potential, Comprehensibility, Challenge and Uncertainty 8 Visual Composition and Stylistic Attributes Symmetry, Balance/Harmony, Photographic Composition, Naturalness and Realism 9 Social Interestingness Popularity and Virality 10 Creativity Creativity 11 Humor Humor, Irony and Sarcasm 12 Urban Perception Urban Interestingness 13 Saliency Saliency and Attention to different factors, including different experimental setups in computer vision tasks, different demographic spread in user studies, differences in understanding the analyzed concept, its definition, and scope, or merely different preferences for the chosen annotators. Interestingness. Berlyne [9] theorizes interest as a primary factor for human motivation and behavior and points out several defining factors of interest [11], such as novelty, in the context of information theory, pointing out that interest arises when new information is compared already existing information by human subjects. More to the point, Chamaret et al. [20] define visual interest as the power of a visual sample to induce interest in a viewer. Furthermore, Silvia et al. [153] relate interest to learning and the will to explore. Similarly, Hidi and Anderson [87] propose that personal

7 Table 2.2: Taxonomy. List of concepts, grouped by positive, negative or unexplored correlation with interestingness as presented in Constantin et al. [33]. Correlations are studied from a physiological or cognitive (t), user studies-based (u) or computational (c) perspective. Controversies are marked with \*.

Positively correlated Negatively correlated Unexplored • Valence(u,t)\* [74] • Arousal(u,c) [160] • Aesthetic Value(u,t,c)\* [90] • Novelty(u,t,c) [74] • Unusualness(c) [187] • Unexpectedness(t) [124] • Complexity(u,t,c) [153] • Coping potential(u,t)\* [155] • Uncertainty(t) [10] • Balance/Harmony(u,c) [96] • Naturalness(u) [79] • Photo Composition(c) [96] • Humor(t,c) [96] • Urban interestingness(u) [141] • Saliency(u) [54] • Attention(t) [12] • Popularity(u,c)\* [77] • Valence(u,t)\* [173] • Boredom(u,t) [61] • Aesthetic Value(t)\* [146] • Memorability(u) [93] • Coping potential(u)\* [160] • Challenge(u)\* [21] • Virality(u,c) [50] • Popularity(u,t)\* [90] • Familiarity(u) [23] • Dominance • Cuteness • Originality • Distinctiveness • Comprehensibility • Symmetry • Realism • Irony, Sarcasm • Creativity • Urban Perception preferences may be less critical in inducing interest in a person than the appeal of the activity or learning task being performed. Aesthetic value. Aesthetics is mainly defined as a branch of philosophy [186], that studies the appeal and beauty of natural scenes and artistic compositions. In several user studies, authors often tend to use “pleasantness” as a descriptor or synonym of aesthetics [74, 160]. Memorability is defined as an intrinsic propriety of visual samples [92], that measures how likely subjects are to remember the images and videos that are presented to them. Some authors use short-term and long-term memorability [48, 47] separation in describing this visual propriety, thus recognizing that, while a video can be memorable for a short period (several minutes or hours), it can be forgotten in the long run (after several days). Violence. While the concept of violence may seem less subjective than others, studies have shown that human annotators do not necessarily agree on whether a visual sample is violent or not. Several studies have used more than one definition of violence, including during the MediaEval 1 Violent Scenes Detection task [46], where authors proposed an “objective” definition (“physical violence or accident resulting in human injury or pain”) and a “subjective” definition (where violence is defined as images “which one would not let an eight years old child see, because they contain physical violence”). Affective value and emotions. The affective value of media items is defined as their ability to induce a set of emotional responses in viewers [18]. From one perspective, they can be described in a mathematical 2D or 3D space, according to the valence-arousal-dominance axes (or only valence and arousal). The VAD space attempts to map all human emotions on these three axes,

corresponding to pleasure-displeasure measuring the valence or pleasantness of the emotion, arousal-nonarousal measuring the intensity of the emotion, and dominance-submissiveness measuring the controlling nature of the emotion [118]. From another perspective, emotions can be described in a categorical space, where a set of basic emotions are identified and defined. Ekman [53] identifies a set of 6 basic emotions: "anger, disgust, fear, joy, sadness and surprise", while Plutchik [132] considers 8 bipolar emotions: "anger-fear, joy-sadness, anticipation-surprise and trust-disgust".

<http://www.multimediaeval.org/> 2.2 Human understanding of the subjective properties of multimedia data

The human understanding of these concepts is extensively studied in psychological and philosophical works. The most important discussion topics here are related to how media samples influence human perception and what underlying factors create that influence. Interestingness. Berlyne [10, 11] identified a series of factors that influence general interest, including conflict, complexity, novelty, and uncertainty. However, these relationships are more complex, as proven in [155], as relationships may not be linear. For example, while novel information is important in inducing interest, subjects may lose interest if that information is too complex to understand. Novelty is also proposed as an important factor for interest in [169, 153]. Hidi and Anderson [87] also show that powerful emotional content has the ability to induce interest, analyzing sexual and violent content as examples. Other works look at the functional benefits brought by interest. Izard and Ackerman [95] conclude that interest is a motivational evolutionary trait, as it allows humans *to explore, learn, and engage with* their *environment*. It is presented *as* one of the main factors contributing to individual adaptation to the environment, survival, and development. [154, 63] also conclude that with the help of interest, in the long run, people are attracted to new possibilities and experiences. Finally, from a physiological point of view, Hess and Polt [86] show that interesting activities influence and are correlated with eye movements and pupil dilation. Aesthetic value. While the aesthetic value of a picture may seem very subjective, theories suggest that some common baseline can be established that most people would agree with. Reber et al. [133] propose "goodness of form, symmetry and figure-ground contrast" as qualities necessary for an item to be deemed beautiful, as such properties would allow human assessors to process that object correctly. Furthermore, with regards to visual beauty in general and to image beauty in particular, Datta et al. [38] propose that, while a normal viewer may be interested in the general effect that an image has ("how soothing a picture is to the eyes"), professional artists may be inclined to analyze other aspects, such as meaning, the use of colors and contours, sharpness and the general "rules of photography". It is also interesting to note that in some works, interest and aesthetics have been studied as correlated concepts or, in the least, concepts that can derive from each other. This idea is best exemplified by Schmidhuber [146], who proposes that "interestingness is the first derivative of beauty: What is beautiful is not necessarily interesting. A beautiful thing is interesting only as long as it is new, that is, as long as the algorithmic regularity that makes it simple has not yet been fully assimilated by the adaptive observer who is still learning to compress the data better". Memorability. Early studies [151] regarding the memorability of images show an impressive human capacity for remembering images in the long term, even when compared with the storage capacity for other objects or concepts such as words or sentences. Furthermore, Brady et al. [13] proved that humans do not simply memorize the general scene in an image (that the authors called "gist"), but are able to encode details correctly and remember even small details and differences between images. According to [136], this capacity is further increased when subjects make a conscientious effort



to memorize the images shown to them. Several other works [92, 93, 17] show memorability to be an “intrinsic propriety of images” and a dependence of memorability on the setting and objects in an image. For example, images containing people seem to be the easiest to memorize, while nature landscapes seem to be the hardest. Furthermore, time plays an important factor in memorability, whether we are talking about the difference between short-term and long-term memory, as presented in [26] or about the time a subject spends looking at an image [136].

11 Violence represents a diverse subject, given its inherent subjectivity and its perception, that can be different from society to society and from generation to generation. Ardent [5] studies violence from a modern perspective, going through some of its possible factors such as “power, strength, force, and authority”. At the same time, Galtung [66] attempts to study it from a cultural perspective, noticing the intra-cultural difference of perception of violence. While these works represent politically-oriented studies on violence, numerous other researchers studied the impact of visual violence in TV, movies, and media. Culbert [35] studies two televised violent events in 1968 (the Tet Offensive and Chicago’s DNC) and analyses the way these events changed public opinion or affected viewers at that time. The same impact is studied in [91], where the authors talk about short- and long-term effects of over-exposure to violence, including the desensitization of casual viewers and the effects on children and young adults.

Affective value and emotions. Psychology is the domain that started to look at the impact of emotional images on human reactions. Valdez and Mehrabian [174] explored the link between colors and the emotions that images are supposed to infer. From a different perspective, Chen and Sun [22] studied the mechanisms that allow emotions conveyed by multimedia teaching material to affect students and their learning performance. Furthermore, understanding emotions may prove useful for understanding other concepts. For example, boredom is used as an antonym of interestingness in [154, 61], and, while not precisely direct opposites, interestingness pushes human subjects towards activity and boredom pushes humans towards inactivity and limits the maximum level of interest that can be achieved [11]. Regarding the 3D representation of the VAD space, generally, valence and arousal are considered to be the most important and most frequently researched [160]; however, some scientists propose a fourth additional dimension, namely “novelty” or “unpredictability” [62], as 12 the addition of this dimension would better represent certain corresponding emotions from the discrete emotional space, the most relevant of them being “surprise”.

### 2.3 Datasets and user studies

Gathering an adequate dataset represents one of the most critical preliminary aspects of creating automated systems to predict such subjective proprieties. While datasets are essential in general for machine learning tasks, in this particular case, some additional matters must be taken into account, such as the difference in opinion between annotators, given the inherently subjective nature of the analyzed multimedia data. Table 2.3 summarizes the primary datasets used for predicting the concepts [defined in the previous section](#), indicating [the type of](#) media files included in the dataset (image or video), the list of annotated concepts, and the types of annotators. While most of the datasets are annotated by human assessors, either through crowdsourcing or through the use of “trusted” annotators that know the task well and are, in some cases, monitored continuously by super-users or master annotators, other approaches involve extracting their annotations from social media platforms directly. In this latter case, standing out are datasets that incorporate information from Flickr or Photo.net

3, platforms that already provide some types of automatic or user-based annotations. Interestingly, researchers also create an extensive collection of datasets that annotate more than one concept, an

approach that may be very useful for predicting subjective concepts in the context of integrating covariates in the feature set and for analyzing inter-concept correlations. The visInterest [160] dataset, composed of 1,005 images, is collected for the study of interestingness and some of its components theorized in physiological works, such as coping potential, complexity, arousal, etc. Another example from this category is represented by two datasets created in the 3https://www.photo.net/ 13 Table 2.3: A list of relevant datasets for the subjective concepts we analyze in this thesis. We present the types of media annotated in the dataset (image or videos), the annotations provided by the authors and the types of annotators used: c - annotations collected through crowdsourcing, t - trusted annotators, w - annotations performed via social media websites, u - unknown annotation sources. Media type Dataset Annotations Annotators Scene categories, interestingness [74] interestingness c Memorability, interestingness [74] interestingness, memorability, aesthetics, unusualness, etc. c visInterest [160] interestingness, arousal, quality, coping potential, complexity, naturalness, familiarity, pleasantness c LaMem [102] memorability c image IAPS [109] VAD space and [amusement](#), [anger](#), [awe](#), fear, [contentment](#), [disgust](#), [excitement](#), [sadness](#) t Abstract paintings [116] [amusement](#), [anger](#), [awe](#), fear, [contentment](#), [disgust](#), [excitement](#), [sadness](#) c Emotion6 [130] VA, anger, disgust, fear, joy, sadness, surprise, neutral c 15K Flickr [144] beauty c Photo.net [38] aesthetics, originality w Aesthetics and interestingness [51] aesthetics, social interestiness w AVA [120] aesthetics w image & video MediaEval Predicting Media Interestingness [48, 47] interestingness t Youtube dataset [96] interestingness t gifInterest [77] interest, aesthetics, VA, curiosity c NHK [167] aesthetics u VideoEmotion [97] [anger](#), [anticipation](#), [disgust](#), [fear](#), [joy](#), [sadness](#), [surprise](#), [trust](#) t video LIRIS-ACCEDE 2 GIFGIF [98] amusement, anger, contempt, disgust, embarrassment, fear, guilt, happiness, pleasure, etc c VA, violence, fear c Movie Memorability [25] memorability t Webcam interestingness t MediaEval Predicting Media Memorability [24, 31] memorability t VIF [83] violence c MediaEval Violent Scenes Detection [44, 45, 46, 159, 158] violence t same work [74]. The authors considered two publicly available datasets, one on scene classification [123] composed of 2,688 images and one on memorability [94] composed of 2,222 images, and annotated them with interestingness values. Another dataset annotated with several concepts is gifInterest [77], where the authors create annotations for interestingness, aesthetics, curiosity, and the violence-arousal space. This dataset is composed of 6,119 video samples, encoded as GIFs. For the prediction of media memorability, a large image-based dataset consisting of over 58,000 samples is presented in [102]. For predicting affective content, authors take several types of approaches, given the different ways emotions are interpreted. While most of the datasets provide annotations for the VAD or VA emotional space [109], there are some examples where only categorical emotional space is used [116]. 14 Finally, datasets such as those published during the MediaEval benchmarking competitions, listed in Table 2.3, annotated for interestingness, memorability, and violence, are also of great importance, as they provide participants with not only a dataset of media samples but also with a common evaluation framework, consisting of concept definition and use case, training/testing data splits, metrics and comparison baselines. These datasets also represent some of the largest collections available to date on their specific tasks. Creating a common evaluation framework for specific tasks can be vital for driving the development of computer vision methods forward, as it creates an accurate baseline for comparing the performance of individual methods, algorithms, and data augmentation approaches. The first step and backbone of creating these datasets are represented by the user studies associated with them. Based on the studies and the answers

returned by the annotators, researchers can create accurate ground truth data that represents the targeted concepts. From a behavioral standpoint, researchers can also deduce the way viewers interact with multimedia data and the visual cues used by annotators in making certain decisions. Some of the most important and interesting user studies consider lists of covariates for the targeted concepts and analyze positive and negative correlations between concepts. For example, Soleymani [160] studies the [link between interestingness and several other concepts](#), concluding [that arousal is the most important attribute](#) for creating [interest](#). Simultaneously, the importance of arousal is also backed up by other works, including [59, 77]. As expected, high correlation values are also found for concepts that psychological theories mention as components of interest. Some examples would include novelty [21] and complexity [160, 2]. Coping potential, another one of the theoretical indicators of interest, is also experimented from a personality perspective. More precisely, some works [160] found that coping potential may have an adverse effect of interest for subjects with high openness trait. In contrast, complexity has a more positive effect on the same group, when compared with subjects that displayed opposing personality traits. Other studies deal with multiple concepts. This is often the case for interestingness, memorability, social interestingness (or popularity) and aesthetics. The studies conducted by [93, 74] conclude that interestingness and memorability are negatively correlated. The studies were conducted on a set of 2222 images. Participants are asked to give their opinion on certain aspects of images (i.e., "Is this image interesting?", "Is this image memorable?"). In the final test regarding memorability, image samples are tested, and a distinction is created between "assumed memorability" (i.e., samples that annotators think they will be able to remember) and actual memorability. Interestingly, while actual [memorability is negatively correlated to interestingness](#), assumed [memorability is](#) positively correlated to it, suggesting that human judgment is not adequate for memorability assessment and that memory tests must be performed to ensure a correct ground truth annotation. Another relationship that is often studied is the one between aesthetics, interestingness, and popularity. Studies [90, 74, 93] show that, while visual interestingness and aesthetics are positively correlated, the same is not true for popularity. Conversely, [Gygli and Soleymani](#) [77] find [a positive correlation between](#) popularity [expressed](#) via the number of likes received by an image and visual interestingness expressed by human annotators. The annotation protocol chosen by the authors vary. For example, Hsieh et al. [90] evaluate visual interestingness on a scale of 5 options, starting from "very boring" to "very interesting", while social interestingness is measured by social networking scores provided by the original websites where these photos are hosted.

#### 2.4 Computational approaches

In recent years, computer vision algorithms started to increasingly target tasks that try to predict the affective value of multimedia data. This is an important step forward, but it requires constant collaboration between different branches of science. To understand the affective and subjective properties of images, computer vision scientists need to have access to large quantities of data, which would allow them to develop their methods and ensure their scalability. As we will present in the following chapter, computer vision algorithms that predict such subjective concepts are relatively new. Most of these branches started their development in the last decade, unlike more traditional [computer vision tasks such as character recognition, object detection, and image classification](#). This chapter presents the advances made by computer vision algorithms in predicting these affective concepts.

##### 2.4.1 Interestingness

One of the first attempts at predicting image interestingness is presented in [74]. The authors used three factors in

determining the interestingness score: novelty, aesthetics, and general preferences. The authors predict these sub-concepts via traditional visual features, i.e., a LOF approach for novelty prediction, aesthetic value using features proposed by [38, 116, 101], and finally, general preference determined by a set of GIST, SIFT, and color histogram descriptors in an RBF-kernel SVM. The authors point out that general preference represented the most important feature with regards to interestingness prediction. Another approach is taken by Fan et al. [58], who conclude that dataset fusion is needed in order to obtain the best results. This may indicate that, given the subjective nature of interest, a larger-than-usual number of visual samples are needed to predict interestingness correctly. For video sample prediction, Jiang et al. [96] use a series of visual, audio, and high-level attributes. The authors find that an early fusion of visual and audio features is the optimal approach in their experiments. Jou et al. [98] perform a comparison between sentiment features and a DNN approach, based on C3D [171], finding that sentiment features perform better. Grabner et al. [73] build a system for interestingness prediction in video streams. The authors build a complexity feature based on compressed file size and a novelty feature based on LOF, achieving good results and confirming covariate-based hypothesis theorized during user studies on interestingness. An unsupervised approach is developed in [115], where images are compared via SIFT descriptors with images with comparable subjects taken from Flickr4. Here the authors base their choice of baseline Flickr images on previous findings that conclude [that Flickr users tend to curate their image collections](#) [105]. The MediaEval Predicting Media Interestingness [48, 47] task gave the opportunity to test several systems in the same setup with regards to dataset, training / testing splits and metrics. While many systems were submitted to the benchmarking competition, some of them stand out. Liem et al [114] propose a system that, among other information, includes features that describe the presence of humans in an image, by extracting color and geometrical descriptors for the human faces from images, concluding that many times faces attract attention and interest. Shen et al. [150] use an SVM based training model that integrates deep features extracted from the AlexNet [107] DNN model. For video processing, Ben-Ahmed et al. [8] employed [deep visual and audio features](#) based on [VGG](#) [157] and [SoundNet](#) [6], trained with a sigmoid kernel SVM. Another relevant approach is presented by Parekh et al. [125], who emulate the annotation process, by automatically developing a pairwise comparison between samples based on deep features extracted from the AlexNet DNN 4 <https://www.flickr.com/model>. For a complete overview of the MediaEval Predicting Media Interestingness task, we refer the reader to [29].

#### 2.4.2 Aesthetic value

Several papers base their approach on previous human studies on aesthetics, composition, and general photography rules. Some essential works here include [101, 38, 39, 112]. These authors designed a large set of traditional visual features centered on human perception and that are accurately able to encode some of these principles, such as depth-of-field, rule of the thirds, and "pleasant" hue combinations, object proportions, etc. These rules, taken in their entirety or just as parts of them, are still exploited in this domain [78, 85]. Regarding more modern approaches, CNN-based systems are starting to show promising results and are implemented by several authors [85]. Furthermore, a multi-patch aggregation method, based on Inception-V3 [166] models is proposed by Wang et al. [178], while Xu et al. [184] use a combination of visual features and an attention-based DNN for predicting aesthetic value. It is important to note that aesthetics is being intensely studied alongside other concepts such as [visual interestingness](#) and [social interestingness](#) (or [social network popularity](#)). Several authors show [a positive correlation between visual interestingness and aesthetics](#)

[51, 74, 90] [from a computer vision perspective](#), while social [in-](#)terestingness shows negative or no correlation with aesthetics. This may result from popularity having more to do with the original poster or the current news and internet trends than with the visual quality of the posted images or videos. However, Redi and Merialdo [135] use aesthetic appeal as an indicator of social interestingness using semantic and composition features on a Flickr based dataset.

#### 2.4.3 Memorability

Early methods for memorability prediction [92, 93] merge human studies with computer vision methods for image classification, using conclusions drawn from the former in designing the latter. Based on the conclusion that memorability is influenced by the objects in a scene, Isola et al. [93] create a set of algorithms based on object statistics and scene descriptors and trained an SVR-based model for memorability prediction. Another significant contribution of this work is an estimator of object type importance based on memorability ground-truth value, thus creating a method for understanding why a photo is memorable. Some experiments are also directed towards increasing the memorability of an image by modifying it. Thus, style transfer models are adopted by Siarohin et al. [152], which create modified image-seed pair later scored by a selector module, that internally uses AlexNet [107] and VGG [157]. More modern approaches fully use the power of deep neural networks. For example, visual attention mechanisms and LSTM layers [89] are deployed in a ResNet-based convolutional architecture by Fajtl et al. [57]. For memorability prediction on video samples, Shekhar et al. [149] incorporate a series of deep learning, video semantics, saliency, spatio-temporal, and color features. Interestingly, fMRI data is also tested as a predictor of memorability in [80]. Recent developments are centered around the MediaEval Predicting Media Memorability task [24, 31]. Given the opportunity to test many short- and long-term memorability prediction systems in the same setting, we must address the fact that, while both editions of the task use the same dataset, data splits, and metrics, the latest edition shows significant improvements with regards to results. Thus, given the lack of additional training data, this may indicate that participants' memorability systems are objectively better. With this in mind, two systems stand out. Azcona [et al.](#) [7] [employ a large set of](#) traditional [visual](#) features, captions, [and](#) DNN-based features, trained with SVR and BRR methods, while Reboud et al. [134] combine 20 captions and visual information, using a large collection of training methods. Interestingly, both participants achieved top results by creating some weighted late fusion schemes that combine results extracted from lower performance systems into a supersystem with better performance.

#### 2.4.4 Violence

Many different interpretations of violence exist in datasets targeting this concept, ranging from aggression in a public environment [185] to violence in specific contexts, such as the stands of sporting events [121] or Hollywood movies and web videos [44]. As expected, the majority of approaches for predicting this concept are based on video sample assessment instead of using single image prediction, as violence is an inherently temporal concept. In this context, several works stand out. Giannakopoulos et al. [68] deploy motion-based visual features and audio features in an early fusion scheme that is trained via a kNN binary classifier. Gong et al. [71] use a semi-supervised approach to this problem, based on cross-feature learning. Starting from low-level features, the authors create candidates for violence detection and candidates for violent events, based on several labels such as "screaming", "explosions", "gun-shots", etc., and combine the output of the two candidate systems. Several types of temporal integration has been tested for creating violence detection systems such as STIP and motion SIFT [121], collections of flow-vector magnitudes [83] or LSTM-based deep neural networks [81]. The 2011-2015 MediaEval Violent Scenes Detection [44, 45, 46, 159, 158] task proposes a common dataset



and evaluation protocol for violence prediction methods. During this campaign, some methods stand out as outliers for their given years. For example, [145] employs a series of low-level visual features and audio features, trained in an MLP approach, and [129] use the same types of features trained by a hybrid K2 and Bayesian system. Similar to these two cases, many of the top-performing 21 systems employ multimodal feature fusion or multimodal training systems. Temporal aggregation or encoding of individual features or videos is achieved both by traditional methods [147] and deep learning methods based on LSTM [36, 37].

2.4.5 Affective value and emotions. A large body of literature is dedicated to emotional content prediction. Zhao et al. [188] explore a set of high-level features based on harmony and the proportions in an image, linking the aesthetic appeal of visual samples with the emotions they convey. On the other hand, sentiment features [98] and arousal features based on color analysis [174] are used for deriving interestingness score [77]. Regarding the dimensional (VAD) emotional space, Sartori et al. [143] also investigate a series of color-based features for abstract painting emotions prediction. More modern, DNN-based approaches are also tested in both the VAD setup and for the categorical emotional space. Acar et al. [1] compare convolutional network approaches with low-level audio-visual features, obtaining better results with the neural network models. Peng et al. [130] employed a modified AlexNet architecture for the same task. While other concepts may involve a binary prediction (i.e., interesting vs. non-interesting) or the regression equivalent or one-class regression, the problem is more complex for emotions. Research papers in this domain will have to employ either multi-dimensional regression, thus predicting samples with regards to the VAD space, or multi-class or multi-label classification, for the categorical emotional space. However, some works choose to take into account both approaches. For example, the authors in [119] create a set of novel audio-visual features that can successfully handle both types of tasks.

2.5 Applications Great interest is shown for computer vision algorithms that can accurately predict and measure these concepts in the context of extensive image collections, where human input would be impossible to achieve due to the large amount of data that must be processed. While some systems must deal with the prediction of such concepts (for example, the detection of violent videos and images), others must create recommendation lists or proposals based on the prediction of subjective judgments, and others must modify the media samples so that values for certain concepts are maximized or minimized. Considering the impact of human cognitive processes on perception, reasoning, attention, and memorization process [55, 156] the importance of developing machine learning techniques for predicting the effect that media items have on the cognitive process. Image collection and video summarization. Interest in this field is constantly growing, and web services that deal with the storage of huge amounts of personal photos, such as Google Photos 5 must also create automatic tools that can process the albums of users in order to create per-album or annual suggestions with regards to the best pictures in those collections. Such a feature represents one of the many ways websites can increase user engagement and loyalty. On the other hand, large videos can also be summarized in order to artificially create "trailers" or advertisements for those videos, based on several aspects. Thus, current works show a tendency of creating video summaries by measuring emotional content [183], interestingness [75, 76] and memorability [25]. Video summarization is very important for the ever-growing number of video and movie hosting services, as it would allow viewers to quickly and efficiently assess video samples and view them according to their personal preferences. 5<https://photos.google.com/> 23 Media recommendation. While the movie recommendation literature was dominated by traditional approaches, based on past

user activity and similarities between user voting preferences, some recent works are starting to incorporate visual and audio movie analysis in their algorithms [42], and make recommendations based on the audio-visual similarity of media items. Other approaches to movie recommendations also target more subjective matters, such as using image aesthetics for creating video features [40]. Perhaps a better-known application from this domain is the Flickr Interestingness API 6, which recommends multimedia items to users based on a measure of social interestingness. Aesthetics-based recommendation systems for image collections are also proposed by Schifanella et al. [144]. Advertising systems. Both traditional and online ads can benefit from introducing methods that can predict the positive or negative impact that ads have on viewers. Recent findings show that dissimilarities between the general viewer mood and the emotional message contained in the ads create the perception that ads are inherently “bad” or “annoying” 7. Recent studies support these findings, as informativity and creativity, along with empathy, are considered factors for a positive response to advertising [110]. Education. Interest is considered to positively affect the educational process, as it would help students better process the information given to them [88]. Though this may seem obvious, some authors suggest that, in their current form, some traditional learning material and textbooks do not rely on using features that would be able to capture and hold attention [4]. Therefore, such measures of interest and other metrics related to concepts like memorability and creativity need to be introduced in the education environment. Multimedia materials could be selected based on such measures and used as learning tools, considering that interest creates motivation, willingness, and energy for learning and can guide career choices [82]. 6<https://www.flickr.com/explore/interesting/> 7<http://www.tronviggroup.com/empathy-in-advertising>

## 2.6 Conclusions

In this chapter, we presented the main theoretical aspects regarding the analysis of the visual impact of multimedia data. We presented the motivations behind the need for a close collaboration between scientists from different fields of study. We have also given some examples from the [current state-of-the-art](#) literature that show [the](#) advantages [of this](#) collaboration. We presented the definitions and some taxonomies for the concepts that will be used throughout this thesis, namely interestingness, aesthetics, memorability, violence and affective value and emotions. We analyzed the [state-of-the-art](#) advances published [in the](#) current [literature](#) regarding [the](#) human understanding of these properties, datasets, user studies, and computational approaches. We also analyzed the subjective nature of these concepts and presented their current or future applications.

## Chapter 3 Personal contributions

### 3.1 Datasets and evaluation

In this chapter, we will present our contribution to the creation of several publicly available datasets, including the following: (i) Interestingness10k [29] 1, designed [for the prediction of image and video interestingness](#); (ii) VSD96 [34], a video dataset for violent scenes detection; (iii) the MediaEval 2019 Predicting Media Memorability [31] a dataset composed of short videos that are annotated [with short-term and long-term memorability](#) values; and finally (iv) the MMTF-14k [41], a dataset for movie recommendation.

#### 3.1.1 Interestingness prediction

The Interestingness10k [29] dataset is a publicly available 2 dataset and a common evaluation framework, designed [for the prediction of image and video interestingness](#). This dataset was tested and validated during the 2016 3 and 2017 4 editions of the MediaEval Predicting Media Interestingness tasks. My main contributions to this 1Paper under major review 2[https://www.interdigital.com/data\\_sets/intrestingness-dataset](https://www.interdigital.com/data_sets/intrestingness-dataset). 3<http://www.multimediaeval.org/mediaeval2016/> 4<http://www.multimediaeval.org/mediaeval2017/> Table 3.1: The Interestingness10k dataset. In this table we present the composition of the image and video

subsets, for both years, 2016 and 2017. Devset represents the development data, while testset represents testing data. Shown here are the number of movies, samples, average duration in seconds for the samples in the video subtask, and the number of interesting samples.

Year	Dataset	# movies	# samples	avg. duration (s)	# interesting
2016	Image	52	5054	1.06	420
2017	Image	78	7396	1.05	646
2016	Video	26	2342	2.14	261
2017	Video	26	2342	2.14	261

dataset are represented by: (i) analyzing the overall performance of the systems submitted to the MediaEval task; (ii) analyzing the influence of features on the prediction models used during the MediaEval competition; (iii) analyzing the generalization capabilities of prediction models on our data; (iv) creating a set of recommendations with regards to system performance; (v) participating in the annotation process.

**Dataset description** This dataset is created according to a Video on Demand use case scenario, employed at Technicolor 5, where participants are asked to create systems that would accurately select images or videos that would create more viewer interest in the source movie [48]. Image and video samples in this dataset are [extracted from Creative Commons 6 licensed Hollywood-like movie trailers](#) and segments, thus creating a publicly available set of visual data. The data is divided into image interestingness and video interestingness prediction subsets. The first step in creating this data is the extraction of video shots from complete movies, separated by camera fade-outs. While the image subset is populated with middle keyframes extracted from those shots, the video subset is populated with the shots themselves. Some general statistics regarding this dataset, presenting both 2016 and 2017 versions, are available in Table 3.1. The dataset evolved from 5,054 devset samples extracted from 52 movies in 2016, to 7,396 extracted from 78 movies in 2017, considering that the 2017 devset data is composed of the previous year's devset and testset combined. On the testset, in 2016, 26 movies were used, accounting for 2,342 samples, and the same number of movies was used in 2017, generating 2,192 samples. For 2017 an additional four full movies are used for enhancing the testset and test system generalization on longer excerpts (the average duration of full movie shots is 11.4 seconds, compared with 2.14 for regular samples). Finally, with regards to system evaluation, two different metrics were chosen for the two versions of the task. For 2016, the overall MAP performance is computed, while for 2017 participant's systems are ranked according to MAP@10. Overall system performance For the overall system performance analysis, we gathered runs submitted to the MediaEval 2016 and 2017 tasks, and we analyzed the trends and improvements implemented by participants. A boxplot representation of these results is presented in Figure 3.1. In this visual representation, we also include systems developed outside of the MediaEval competition, in [state-of-the-art papers, that use the same rules and validation principles as the ones used during the competition](#). In order to allow an overall comparison between the two years of the competition, we also provide MAP scores for the 2017 edition of the tasks. Also, for each year, three human annotator runs are represented with red dots, representing the prediction performance of humans on this dataset. While human results are above the presented systems, they still do not achieve very high results, further indicating the proposed task's subjectivity.

Year	Dataset	mAP	mAP@10
2016	Image	0.7	0.6
2016	Video	0.5	0.4
2017	Image	0.3	0.2
2017	Video	0.1	0

Figure 3.1: [Boxplot representation of the overall system performance](#). Data is presented as follows: interquartile range (IQR) 50%, median values (red line), [lower and upper adjacent values calculated as  \$Q1 - 1.5 \times IQR\$  and  \$Q3 + 1.5 \times IQR\$  respectively](#). For

reference, [the](#) performance of 3 human annotators is represented with red dots. Regarding overall system performance, the first observation is that no systems represent either positive or negative outliers. Another clear observation is that 2017 systems performed better, with regards to MAP, than 2016 systems, indicating that a larger training set and better, more interest-oriented systems improve the overall results. In the case of the image subtask this improvement is 25.75%, from a MAP value of 0.2485 [30] to 0.3125 [125]. For the video subtask, the improvement is 22.75%, going up from a MAP value of 0.1815 [3], to 0.2228 [180]. Another critical point is that, in general, system performance for video samples is worse than that of image samples, indicating that better methods are needed for video interestingness prediction.

0.32 0.30 0.28 0.26 P0.24 m0.22 A 0.20 0.18 0.16 0.14 0.12 2016.Image  
 2017.Image visual audio deep [audio+deep](#) [audio+deep+](#) text [deep+](#) concepts motion [visual+audio](#)  
[visual+](#) deep [visual+audio+deep](#) concepts [visual+](#) deep+concepts visual+motion motion+audio  
 visual+motion+deep deep+text visual+concepts deep+text+concepts 0.23 0.22 0.21 0.20 P0.19 m0.18 A  
 0.17 0.16 0.15 0.14 0.13 2016.Video 2017.Video 0.16 0.14 0.12 100.10 P@ m0.08 A 0.06 0.04 0.02 0  
 2017.Image 2017.Video

Figure 3.2: Analysis of the employed features: Year.Type represents the year of the data (2016 or 2017) and its type (Image or Video). Official metrics for 2016 data is mAP and for 2017 is mAP@10. For comparison, we also provide mAP for 2017. We represent both the participating systems from MediaEval benchmark as well as state-of-the-art approaches from literature (marked with a red circle). Feature-level analysis Our analysis of the content descriptors employed by the systems submitted to this task attempts to bring to light the contributions of certain types of descriptors and, if possible, make recommendations with regards to the approaches that are best suited for interestingness prediction. We identified six main feature types that were employed by participants, as follows: traditional visual features, audio, motion, deep learning-based features, conceptual and textual. Of course, many systems use not one, but rather a combination of these types of descriptors, generating 18 employed combinations. Figure 3.2 presents the results of these approaches. We also included systems developed outside the MediaEval competition, as well as MAP performance for both years, allowing for better comparisons. More precisely, our analysis shows that, with regards to image interestingness, systems that use deep features perform, on average, better than others, with an average MAP of 0.2297, while for video interestingness, traditional visual features

31 0.32 0.3 0.28 0.26 0.24 A0.22 P m 0.2 0.18 0.16 0.14  
 0.12 none pre-trained ne-tuned correlated

Figure 3.3: System performance for concept generalization. Blue dots represent systems used for image interestingness prediction, while red dots represent systems used for videos. perform better, with an average MAP of 0.1798. On the other hand, when analyzing the fusion schemes employed by participants, we found that late fusion systems are the best performers on average, for both the image and video subtasks. This analysis is further presented in Table 3.2.

Table 3.2: Average MAP for the analyzed systems, grouped by employed features and fusion scheme.

Year	Type	Feature	Fusion	avg. mAP	#systems
2016	Image	visual	early fusion	0.2258	34
		deep	early fusion	0.2297	53
		motion	early fusion	0.2053	5
		audio	early fusion	0.2157	11
		text	early fusion	0.2277	38
		concepts	early fusion	0.2260	35
	Video	visual	late fusion	0.2416	18
		deep	late fusion	0.1798	49
		motion	late fusion	0.1776	61
		audio	late fusion	0.1704	14
		text	late fusion	0.1746	23
		concepts	late fusion	0.1721	6
2017	Image	visual	early fusion	0.1767	12
		deep	early fusion	0.1768	51
		motion	early fusion	0.1731	43
	Video	visual	late fusion	0.1878	30
		deep	late fusion	0.1731	43
		motion	late fusion	0.1878	30

Generalization capabilities We analyze three types of system generalization capabilities: (i) concept generalization, where we analyze the correlations between other concepts and interestingness, (ii) image-to-video generalization, where we test whether systems that predicting image interestingness can represent capable video interestingness predictors and finally 32 (iii) short-vs-long video generalization, where we

compare testset performance on short and long videos. For concept generalization, we theorized four types of systems, as shown in Figure 3.3. Pre-trained systems represent methods that are pre-trained on unrelated data, such as general image or action classification, fine-tuned systems represent methods that initially trained on general classification tasks and then re-trained on Interestingness10k, correlated systems represent methods that use data from other correlated concepts like emotional content or memorability prediction, and finally, none represent systems that use none of these generalization schemes, thus being trained solely on Interestingness10k data. The primary observation in this analysis is that, for the image prediction subtask, pre-trained systems significantly outperform other types of systems, with an average MAP of 0.2405, compared to 0.2208 for systems with no generalization. Unfortunately, no such statistical relevance is found for the video prediction systems. During this analysis, we identified some datasets and models that use positively or negatively correlated concepts and that are used in various system training stages. Some examples would include the methods of Shen et al. [150], that use a dataset of 0.2 million images extracted from Flickr, according to their social interestingness API 7 in the pre-training stage, or Erdogan et al. [56] that extracts the fully connected weights of the MemNet model [102]. With regards to image-to-video generalization, we analyze systems that use the same training schemes for predicting [both image and video interestingness](#). This includes using [the](#) same set of features, training model and architectures, and pre- and post-processing methods. We also incorporate video prediction systems that employ a simple statistical approach (such as averaging) when transforming frame-level features to a global video descriptor. While only ten systems fall into this category, the correlation between image MAP performance and video MAP performance for 7 <https://www.flickr.com/explore/interesting/> 33 those systems, calculated via Pearson's Correlation Coefficient, is 0.546, indicating that, although not a strict statistical proof, adapting image predictors to videos may represent a good starting point. Finally, with regards to short-vs-long generalization capabilities, we separate the short testing samples from the long testing samples in the 2017 edition of the task and calculate the average MAP across these two sets. Results show an average MAP@10 of 0.0562 for short videos and 0.0751 for long videos. We attribute this to the video length difference, which may create a better differentiation between interesting and noninteresting samples. Recommendations with regards to system performance Finally, we drafted a set of important observations and recommendations regarding the construction of an interestingness prediction system. These would include: - deep features (for images) and traditional visual features (for videos) perform better than other types of descriptors; - late fusion systems represent an obvious advantage when compared with systems that employ early or no fusion, this observation being also supported by our proposed DNN-based ensembling model; - systems that use more than one type of classifier or regressor tend to outperform single-classifier systems; - more modern DNN approaches, like GSM-InceptionV3 [163], can have good performances, however they do not surpass the current state-of-the-art; - upsampling can have a positive effect on system performance, as shown in [150]; - system performance may benefit from pre-training on external data [176].

### 3.1.2 Violence prediction

The VSD96 dataset [34] is a publicly available dataset and a common evaluation framework designed for the detection of violent scenes in Hollywood-like and YouTube movies. This dataset is validated during the 2011 - 2015 editions of the MediaEval Violent Scenes Detection tasks. My [main contributions](#) to [this](#) dataset [are as follows](#): (i) [an](#) overall [analysis](#) of systems [that](#) use this dataset, and (ii) an analysis of the types of features employed for violence prediction. Dataset description



An overview of the [dataset is presented in Table 3.3](#). Overall, [the dataset](#) comprises annotated data from 31 full Hollywood movies, 86 YouTube videos, and 199 Hollywood-like movie clips. Several types of annotations are utilized, varying across the different editions of the MediaEval task. For 2011, 2012, and 2013, videos are segmented at shot level, via a shot boundary detector algorithm, and annotations are performed per individual shot. For 2012, 2013, and 2014, we also provide annotations at segment level, containing a starting and an ending frame number per violent segment. Finally, for 2015 annotations are done at the video clip level. Another level of annotations is represented by the definition of violence used by the annotators: (i) an objective definition, i.e., annotators are asked to determine the videos that show “physical violence or accident resulting in human injury or pain”, and (ii) a subjective definition, i.e., a video that “one would not let an 8-year old child see in a movie because it contains physical violence”. Furthermore, several metrics are used for the different versions of this dataset: (i) Cost metric for 2011, (ii) MAP@100 for 2012 and 2013, (iii) MAP2014 for 2014, and (iv) MAP for 2015. 8Data for 2011-2014 available at: [https://www.interdigital.com/data\\_sets/violent-scenes-dataset](https://www.interdigital.com/data_sets/violent-scenes-dataset) 9Data for 2015 available at: <http://liris-accede.ec-lyon.fr/> 10www.youtube.com 11http://www.multimediaeval.org/ Table 3.3: The VSD96 dataset. We indicate the types of movie sources used (Hollywood, YouTube or Hollywood-like) year of the task (2011-2015), number of source movies, their total duration in minutes, number of segments extracted from the movies and the percentage of violent 2014 Movie types 2015 content. 2013 2012 2011 # movies duration (m) # segm % violence Hollywood movies dev dev test dev test dev test 12 3 3 7 1397 318 404 885 21617 4500 6570 11245 13.25 19.91 9.84 12.86 test 7 833 359 16.26 YouTube videos test gen 86 157 86 44.47 Hollywood-like dev 100 1014 6144 4.42 movie clips test 99 784 4756 4.90 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0 MAP Cost MAP MAP@100 MAP MAP@100 MAP MAP@100 MAP MAP@100 MAP MAP2014 MAP MAP2014 MAP VSD2011H-obj.shot VSD2012H-obj.shot VSD2013H-obj.shot VSD2013H-obj.seg. VSD2013H-subj.shot VSD2014H-subj.seg VSD2014YT-subj.seg VSD2015H-subj.clip

Figure 3.4: Overall performance representation. We present system performances per competition, per task, using both the original metric used during the MediaEval competition and a MAP performance, in order to allow for comparisons between editions. Boxplots are created as follows: interquartile range (IQR) 50%, median values (red line), [lower and upper](#) adjacent values [calculated as  \$Q1 - 1.5 \times IQR\$  and  \$Q3 + 1.5 \times IQR\$  respectively](#). Overall system performance Our analysis of general system performance is presented in Figure 3.4. Some improvements are evident in this analysis. For example, given the objective (denoted obj) definition of violence, from 2011 to 2013, MAP performance has increased, reaching 0.51 for shot-level violence prediction. An analysis based on the definition of violence can be performed, especially for the 2013 data, where both definitions are used on the same set of development and training data. Here we can notice that for systems that used shots as inputs, higher MAP and MAP@10 values are attained for the subjective definition of violence (denoted subj). We attribute this to better-balanced data, as there are more subjective violent samples than objective ones, i.e., 20.24% 36 compared with 10.49%, in the training set. Improvements have also been recorded in the 2014 version of the task, where, for segment-level prediction, a MAP value of 0.7 is attained. Also encouraging are the good results recorded on the YouTube generalization dataset (denoted YT). While systems for 2014 are trained on Hollywood (denoted H) movies, they are still capable of detecting violence in the generalization tests performed on YouTube data, indicating that systems are well trained and could perform well even in a more general understanding of violence. Also, for YouTube testing data, the class

imbalance problem is significantly lower than Hollywood data, with 44.4% of this data being annotated as violent. Finally, a significant decrease in performance is registered in 2015, with a maximum MAP of 0.29. However, this may be explained as participants' systems to the 2015 task are required to predict both violence and emotional content in VA space. Feature-level analysis Our analysis of the employed features shows that several types of descriptors are used in the composition of systems. These are: (i) traditional visual features, (ii) audio features, (iii) conceptual features, (iv) deep learning features. Participants also employ combinations of these four features, totaling up to 12 types of single- and multi-modality types of features. While some single modality systems, such as Dai et al. [36], achieve good results by using just traditional video features, with a MAP of 0.706 on VSD2014H-subj.seg, or Tan et al. [168] that uses an extended set of conceptual features, achieving a MAP of 0.675 and 0.674 on VSD2013H-shot.seg, multimodal systems are better performers. On average, single modality systems achieve an average MAP of 0.208, while systems that employ multiple modalities have average MAP results of 0.313, which represents a significant improvement. Furthermore, with regards to multimodal systems, four categories stand out, obtaining top results in certain subtasks: (i) visual and audio [72, 49], (ii) audio and conceptual [128], (iii) 37 visual, audio and conceptual [127, 145], and finally (iv) visual, audio and deep [37]. Furthermore, with regards to late fusion, ensembling systems achieve an average MAP of 0.343, thus further suggesting the advantages of late fusion schemes.

### 3.1.3 Memorability prediction

The MediaEval 2019 12 Predicting Media Interestingness dataset [31], is a dataset validated during the 2019 edition of the MediaEval Benchmarking Initiative. This task requires participants to accurately predict the short- and long-term memorability for video samples. For this dataset, my main contribution is leading the organization team during the MediaEval competition. Dataset description The dataset is annotated with short- and long-term memorability ground-truth values, corresponding to human annotators' ability to remember whether they previously saw a video or not. For the short-term memorability, videos were repeated in the same annotation cycle, only tens of minutes away from their first appearance, while the long-term memorability is tested by the same annotators, 24-72 hours after the short-term cycle. The dataset is composed of 10,000 short soundless videos, with an average length of 7 seconds. The data is split into 80% development set data, corresponding to the videos that participants must use to develop their systems and 20% testing data. For this task, the official metric is Spearman's rank correlation.

MediaEval 2019 Predicting Media Interestingness During this edition of the Predicting Media Interestingness task, eight teams participated in both the short- and long-term tasks. Results are encouraging, as shown in Table 3.4. The best performing systems are developed by Azcona et al. [7], with a correlation of 0.528 on the short-term prediction task and Reboud et al. [134], with a correlation of 0.277 on the long-term task. Considering the improvement in top results recorded in 2019 compared with 2018 and the fact that every system presented at this edition performs above 2018's average correlation score, we consider 12

<http://www.multimediaeval.org/mediaeval2019/> 39 Table 3.4: Results during the 2019 Predicting Media Interestingness task. For comparison, we also present the best and average results from the 2018 edition of this task.

Team	Best short-term result	Best long-term result
Insight@DCU [7]	0.528	0.27
MeMAD [134]	0.522	0.277
Best 2018	0.497	0.257
UPB-L2S [32]	0.477	0.232
RUC [181]	0.472	0.216
EssexHubTV [111]	0.467	0.203
TCNJ-CS [177]	0.455	0.218
HCMUS [172]	0.445	0.208
GIBIS [142]	0.438	0.199
Average 2018	0.359	0.173

this edition to be a success, driving forward the computational understanding of media memorability. Some general trends and processing methods that

improve system performance are: using ensemble or late fusion systems, deep features, and feature dimensionality reduction.

### 3.1.4 Content recommendation

The MMTF-14K [41] is a publicly available dataset that creates a collection of data for Hollywood movie trailer recommendation systems. While most recommender systems and datasets base their decisions on metadata-like features, consisting of user ratings, movie genres, and other related descriptors, this dataset also provides audio and visual features that can help the recommendation process, creating a multimodal decision system. My main contribution to this dataset is represented by the computation of visual deep learning-based features and visual aesthetic features.

#### Dataset description

The dataset is based on ratings and movies extracted from the popular MovieLens 14 dataset. User ratings are expressed on a scale of 1 to 5 stars, while the entire dataset is composed of 13,623 movie trailers, for which approximately 138 thousand users created over 12 million individual ratings. This dataset also provides metadata features that describe the movie's genre and user-generated tags, audio features represented by BLF and i-vector features, and finally, [a set of visual features extracted from the AlexNet \[107\] DNN](#) and aesthetic visual features as presented in [38, 101, 112, 78]. The MRR, MAP, and R metrics are calculated at different cutoff points (i.e., 4 and 10). These features, along with their early fusion combinations, constitute a baseline collection of methods that can be used as a baseline for comparing future methods employed by researchers who want to use this dataset.

#### Visual features

The aesthetic visual features are a collection of descriptors, aggregated by Haas et al. [78] and developed in several works on image aesthetics [38, 101, 112]. This set of 26 features targets image aesthetics from three different perspectives: color-, texture- and object-based aesthetics. We present three possible feature early fusion combinations: individual features, features grouped according to the three perspectives, and a fusion scheme containing all the features. For the deep AlexNet [107] features, we [extract the output values of the fully connected fc7 layer](#). We provide video-level aggregation for both these features and their early fusion combinations starting from frame-level feature extraction via simple statistical aggregation, i.e., [average, median, average + variance, and median + median absolute deviation](#).

### 3.2 Predicting media interestingness

#### 3.2.1 Introduction

In this chapter, we present the contributions concerning the prediction of media interestingness. We propose implementing SVM-based learning systems that use several visual features [27] as well as learning systems based on the use of aesthetic features and late fusion [28, 30]. The main contributions consist of applying a set of traditional visual features and a set of finely-grained aesthetic features to the domain of visual interestingness prediction and applying late fusion schemes in order to improve final system performance. Experiments with these approaches are carried out in the context of two consecutive benchmarking campaigns that provide two incremental datasets for [both image and video interestingness prediction](#), namely [the MediaEval 2016 \[48\] and 2017 \[47\] Predicting Media Interestingness tasks](#).

#### 3.2.2 SVM-based learning systems

##### Motivation

As summarized in our literature survey paper [33], the concept of visual interestingness in highly subjective. Current state-of-the-art literature shows several concepts [to be both positively and negatively correlated with](#) interest. For example, while Gygli et al. [74] show valence as a positive contributor to interest, Turner and Silvia [173] show it to be negative, other examples including popularity [77, 90] and coping potential [155, 160]. While many factors can create and increase this type of subjectivity, one of the most important factors is the human annotators' personal preferences and opinions. Considering these factors, we decide to use a set of features that are traditionally used in

the creation of image descriptors, thus testing a diverse baseline for this task. We present this approach in our paper [27].

Media samples Feature extraction Feature set Feature fusion SVM Results [Figure 3. 5: The diagram of the proposed SVM-based method. The three main stages \(Feature extraction, Feature fusion and SVM\)](#) are highlighted in blue. Previous work Several works have studied the contribution of visual features to media interestingness prediction. For example, Soleymani [160] uses HOG, LBP, and GIST [123] as visual features, and trained these features using a regression that employs sparse data approximation [122] for image interestingness. Other approaches include using colorfulness [38], arousal values [116], JPEG compression size and edge distribution [101] in [74] and a semantic content detection algorithm based on Fast-RCNN [69] developed in [117]. For video interestingness prediction, Jiang et al [96] use traditional and high-level attribute features, including HSV color histogram, SIFT, GIST, Classemes [170] and style attributes [120]. Gygli and Soleymani [77] also use a set of visual descriptors, while Jou et al [98] implements a set of sentiment-based features. Proposed approach Our approach consists of three phases, as described in Figure 3.5 and is described in [27]. The first stage consists of processing the media samples by employing a set of features extractors, while the second stage consists of applying feature-level fusion. Finally, the last stage consists of using an SVM-based approach for classifying the media samples. A set of seven descriptors are extracted for each media sample. These features include: (i) color histogram calculated in the HSV space (denoted HSVHist), (ii) 44 dense SIFT transform with a 300 words codebook (SIFT), (iii) Local Binary Patterns (LBP), (iv) HoG descriptors calculated over densely sampled patches (HOG), (v) GIST computed with Gabor-like features (GIST), (vi) a couple of features extracted from the FC7 and Prob layers of the AlexNet architecture [107] (ANfc7 and ANprob) and (vii) the color naming histogram proposed in [175], that provides a lower-dimensional space of values for the colors in an image. For image processing, each sample is represented by this collection of feature vectors. In contrast, for video processing, we create a global video-level descriptor by averaging the vectors of all the individual frames. Regarding the feature-level fusion, we choose every combination of two individual features and, starting from that point, combinations of three best performing features, thus creating a total of 39 feature combinations for each subtask. As previously mentioned, the final stage is represented by an SVM-based learning method. To maximize the system's performance, we choose a broad set of experiments and start by implementing polynomial, RBF, and linear kernels. The following SVM parameters are tested for the polynomial kernels in order to optimize the results: - polynomial degree (denoted  $d$ ) with values of 1, 2 and  $3 \times k$ , where  $k \in [1, \dots, 10]$ ; - gamma coefficient (denoted  $\gamma$ ) with values of  $2k$ , where  $k \in [0, \dots, 6]$ ; while for the RBF kernels the following parameters are tested: - cost (denoted  $c$ ) - gamma, both with values of  $2k$ , where  $k \in [-4, \dots, 8]$ .

Experimental setup These methods are tested in the context of [the MediaEval 2016 Predicting Media Interestingness Task](#) [48]. The task defines interestingness in a Video-On-Demand use case, where participants are tasked with selecting images and videos that are 45 most interesting for a "common viewer". This dataset is presented and detailed in Section 3.1.1. Experimental results The experiments are carried out in two stages. While in the initial stage, [using a 10-fold cross-validation](#) method, [we select the best-performing](#) methods with regards to MAP performance on the devset, [in the final stage, we run the best-performing](#) systems [on the](#) testset, thus obtaining [the](#) final system performance. As a general observation, all the best-performing systems use polynomial kernel. Given the limit of five submissions per team for the final testing stage, we start by selecting the best five performers for the image and video

subtasks on the devset, presented in Table 3.5. While for the image subtask, the top system with regards to the MAP metric is a polynomial SVM with  $d = 15$  and  $\gamma = 2$  that uses HSV histogram and GIST features, achieving a MAP score of 0.214, for the video subtask, a polynomial SVM represents the top system with GIST and ANprob features, and  $d = 9$  and  $\gamma = 5$ , achieving a MAP of 0.179. It is interesting to note that systems that include early feature-level fusion outperform single-feature systems. For comparison, in the image subtask, the best-performing single-feature system uses colormnames, resulting in a MAP of 0.195, while for the video subtask, it is represented by a GIST- based system that achieves a MAP of 0.148. In the final stage, we use the top 2 image systems and the top 3 video systems. Finally, the selected systems are run and tested on the testset. Their results are compared with the top performer and average MAP score from the MediaEval 2016 Interestingness task and presented in Table 3.6. As provided by the task organizers, we also present precision values for several cutoff points: 5, 10, 20, and 100. The majority of the systems we submitted present better MAP results on the devset they were trained on, with a single exception represented by the SIFT+ANprob 46

Table 3.5: [Best results on devset for the image and video subtasks](#). We present the subtask (image or video), features that compose the systems, type of SVM employed and results for the Precision, Recall and MAP metrics. Task Feature SVM type ( $d, \gamma$ ) Precision Recall MAP image HSVHist +GIST poly (18, 2) 0.224 0.05 0.214 image SIFT +GIST poly (3, 32) 0.16 0.144 0.211 image HSVHist+SIFT +GIST poly (9, 2) 0.3 0.034 0.197 image colormnames+any poly (3, 2) 0.143 0.128 0.195 image colormnames poly (2, 8) 0.107 0.517 0.195 video GIST+ ANprob poly (9, 4) 0.103 0.083 0.179 video ANfc7 +any poly (3, 4) 0.099 0.095 0.172 video SIFT+ANprob poly (24, 64) 0.087 0.192 0.159 video GIST poly (6, 8) 0.121 0.116 0.148 video SIFT poly (3,64) 0.109 0.059 0.147 run on the video subtask. Overall, for the image subtask, the results were below the average MediaEval values, while for the video subtask, all the runs were over the average MediaEval performance, without reaching the top performance. Considering MAP, the official metric of this task, we achieve the highest performance for the submitted systems with an HSVHist + GIST combination for the image subtask (MAP = 0.1714) and SIFT + ANProb for the video subtask (MAP = 0.1629). Table 3.6: Final results of the selected systems on the testset. The results are compared with the top performer (ME top) and average (ME avg) MAP from the MediaEval interestingness task. Results are also compared with regards to Precision metric (P) at different cutoff values (5, 10, 20, and 100).

Subtask	System	MAP	P@5	P@10	P@20	P@100
image	ME top [114]	0.2336	-	-	-	-
image	ME avg	0.2009	-	-	-	-
image	HSVHist+GIST	0.1714	0.1077	0.1346	0.1423	0.0869
image	SIFT+GIST	0.1398	0.0462	0.0808	0.1000	0.0862
video	ME top [3]	0.1815	-	-	-	-
video	SIFT+ANprob	0.1629	0.1154	0.1500	0.1192	0.0819
video	GIST+ANprob	0.1574	0.0923	0.1269	0.1212	0.0812
video	ANfc7+HSVHist	0.1572	0.1231	0.1000	0.1077	0.0815
video	ME avg	0.1572	-	-	-	-

### 3.2.3 Aesthetic features and late fusion learning systems

#### Motivation

Given the previous results [27] presented at the MediaEval 2016 interestingness task, the need to implement methods that are more tuned for interestingness prediction becomes more apparent. As presented in our literature survey paper, [33], aesthetic appeal and interestingness are quite often studied together. Previous works in psychology [87] and user studies [74] found a positive correlation between aesthetics and interest. While some authors also found negative or low correlations between these two concepts [146], we nonetheless decide to extract a set of aesthetic-based features, developed in [38, 112, 101] and use these features for the prediction of media interestingness. We test this approach on the MediaEval 2016 [48] and 2017 [47] Predicting Media Interestingness Task datasets. We present these



in two of our papers [30, 28]. Previous work Starting from psychological works and user studies from literature, some authors have previously used aesthetic-based computer vision methods for the prediction of media interestingness. However, these approaches usually use few aesthetic cues, under the form of a low-dimensional feature vector. For example, Gygli et al.[74] used an aesthetic descriptor, composed of colorfulness values [38], arousal [116], [complexity based on JPEG size and contrast and edge distribution](#) [101]. Jou [et al.](#) [98] used simple visual features often associated with aesthetics such as brightness and balance, in creating a baseline for comparing their proposed systems. Proposed approach Our approach uses a set of aesthetic feature extractors developed in [38, 112, 101, 78], that are trained using SVM classifiers with polynomial, RBF, and linear kernels. We 48 Media samples Feature extraction Feature set Feature fusion SVM Results Results Results Late fusion [Figure 3. 6: The diagram of the proposed aesthetic-based method. The](#) four main stages (Feature extraction, Feature fusion, SVM and Late fusion) are highlighted in blue. attempt to increase system results by employing two types of fusion experiments: early fusion and late fusion schemes. [A graphical representation of these systems is presented in Figure 3.6.](#) With regards to the aesthetic descriptors, three main groups of features are used in this work, as described in [78]: (i) color-based features, (ii) texture-based features, and (iii) object or segmentation-based features. Some of these are heavily inspired by research conducted in correlated domains, such as color theory, photographic practices, and image composition. The following features are part of the color-based group: - Color values in HSV and HSL space implemented as average over the three space components (denoted HSV, HSL); - Colorfulness implemented as quadratic-form distance and as Earth-Mover distance [101], and as standard deviation [78]; - Hue statistics, according to the findings in [112, 101] that study the importance of hues on human aesthetic perception (HueDesc); - Hue models presence, as Li et al. [112] proposed a set of 9 hue combinations that are more pleasant (HueModel); - Brightness, calculated as brightness contrast across the image according to the methods presented in [112]; 49 - [Average HSV values, based on the Rule of the Thirds](#) from image composition theory, as presented in [38] (aHSVRot); - Average HSL values calculated around the focal point of the image, as presented in [112] (aHSLFocus). We employ the following texture-based features: - Edge, calculated as edge energy as presented in [112, 101] and sum of edges [78]; - Range of textures, calculated at  $3 \times 3$  bounding boxes as presented in [78] (denoted Texture); - [Entropy of the red, green and blue spaces](#), as described in [78] (RGBEntropy); - HSV Wavelet functions, [a three level Daubechies wavelet transform, implemented](#) by [38]; - Low depth of field indicator, as presented in [38] (DoF). Finally, the following object-based features are employed: - Size of the largest 5 segments in an image, as proposed by [38] (denoted LargSegm); - Centroids for the 5 largest segments in an image, as described by [38]; - HSV and brightness average values for the largest 5 segments, as proposed by [38, 112] (HueSegm, SatSegm, ValSegm, BriSegm); - Color model for the largest 5 segments, calculated based on average color spread and complementary colors [38] (ColorSegm); - Coordinates of the largest 3 segments, as presented by [112] (CoordSegm); - [Mass variance and skewness for the largest 3 segments](#) [112] (MassVarSegm, SkewSegm); - Contrast between segments, between the HSV and blur attributes, as described in [112] (ContrastSegm). 50 We use the same early fusion and SVM training parameter variance schemes like those presented in Chapter 3.2.2. Furthermore, in the final step, we deploy a series of traditional statistical late fusion methods to increase system performance. In general, late fusion, or ensemble [learning, is defined as a series of methods that,](#) by combining the classification or regression outputs of

several weaker learning systems, called inducers, can provide a better set of predictions for a given problem. The methods we use in this work are the following: (i) CombSum, (ii) CombMin, (iii) CombMax, (iv) CombMean. The first of these methods consists of summing the prediction outputs of the inducer systems, while CombMin and CombMax take the minimum and the maximum value respectively of the inducer's prediction outputs. The last method consists of a weighted summing of the inducer outputs, according to the formula:  $N \text{ CombMean}(\text{Img}) = \sum_{i=1}^N w_i o_i$  (3.1) where  $N$  represents the number of inducers selected for the experiment,  $o_i$  represents the outputs of individual inducers and  $w_i$  represents the weight applied to each inducer. For our experiments, we choose the following values for  $w_i$ :  $w_i = 2 \text{ran}1k(i)$ . The rank(i) function returns 0 for the best performing inducer, 1 for the second best and so on. Experimental setup Experiments are conducted on the datasets presented at the MediaEval 2016 [48] and 2017 [47] Predicting Media Interestingness Task. While the 2016 version of this dataset was also used for the experiments in Chapter 3.2.2, the 2017 version represents an extension of that dataset, with more samples for the training and testing sets. Also, for the 2016 version of the dataset, we only conducted experiments on the image subtask. Both the datasets are presented and detailed in Section 3.1.1.

51 Table 3.7: Final results of the systems on the MediaEval 2016 image Interestingness testset. These results are compared with the top performer (ME top) and average value (ME avg) presented during the benchmarking competition, according to the official MAP metric. Our top five systems are presented, all of them being late fusion systems, along with the best early fusion and inducer system.

Approach	MAP
Late fusion 0.2485 CombMax (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)	0.2485
Late fusion 0.2451 CombMean (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)	0.2451
Late fusion 0.2448 CombMean (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)	0.2448
CombSum (aHSVWavelet + HueSegm + SatSegm and Late fusion 0.2408 CombMax (aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)	0.2408
Late fusion 0.2403 SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)	0.2403
Early fusion 0.2363 SatSegm + MassVarSegm + SkewSegm	0.2363
ME top [114]	0.2336
Inducer 0.2057 aHSVWavelet or SatSegm	0.2057
ME avg	0.2009

Experimental results Regarding the 2016 version of the dataset, the results of the experiments are presented in [30]. Similar to the experiments presented in Chapter 3.2.2, individual features performed worse than early and late fusion combinations. Table 3.7 presents the results. It is interesting to note that, even though individual inducer systems did not outperform the MediaEval top system, represented by Liem et al [114], they did perform above average. The five best performing inducers are, in order of MAP performance: aHSVWavelet and SatSegm, both with a MAP of 0.2057 and HSV with a MAP of 0.2051. We record an increase in performance when employing early fusion schemes. In this case, early fusion results surpass the top MediaEval performance. The best early fusion schemes are as follows: SatSegm + MassVarSegm + SkewSegm 52 with a MAP of 0.2363, aHSVWavelet + HueSegm + SatSegm MAP performance of 0.2261 and HSL + LargSegm + BrightSegm with a MAP of 0.2232. The late fusion systems achieve even better performances. Table 3.7 presents the top five late fusion systems. In this case, the best performing system is a CombMax late fusion scheme, applied to early fusion feature combinations of aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm, attaining a MAP score of 0.2485. Interestingly, object-based features predominantly produce better results in this experimental setup, whether they are employed in early or late fusion experiments. This

observation may indicate a human annotator preference towards judging the interestingness of images based on the most salient [objects in the scene](#). The [second](#) part of our experiments is carried out on the MediaEval 2017 dataset, both [for the image and video subtasks](#). The [results are](#) presented in Table 3.8. Systems developed for the video subtask perform better, having results above the average MediaEval score. The [best performing system](#) on [the image subtask](#) is a CombMean late fusion scheme that uses aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm early fusion features (MAP@10 = 0.5555), while on the video subtask, it is again a CombMean late fusion scheme that uses LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture early fusion features (MAP = 0.0732). As is the case for the 2016 experiments, the late fusion systems perform better than the early fusion systems, which in turn perform better than the individual inducers. Another general observation is that the RBF kernel shows optimal results for this dataset. Surprisingly, better results are achieved [for the video subtask](#) than [for the image subtask](#). This may result from the training phase, which is adapted to a MAP@10 setting, which perhaps does not allow for a good enough separation and therefore training between the image samples. Finally, we observe that CombMin and CombSum strategies do not improve the results of their individual inducer components.

53 Table 3.8: Final results of the systems submitted at the MediaEval 2017 Interestingness task. These results are compared with the top performer (ME top) and average value (ME avg) presented during the benchmarking competition, according to the official MAP@10 metric and to the additional MAP metric. For the proposed systems, results on the devset are also presented. Subtask

Approach	MAP@10 devset	MAP testset	MAP@10 testset	ME top [131]	ME avg
image CombMean (aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0793	0.1873	0.0555	0.1851	0.0529
CombMean (HSVWavelet + aHSVWavelet + aHSLFocus and HSV + HSL + aHSLFocus and HSV + MassVarSegm)	0.0793	0.1851	0.0529	0.0821	0.1791
CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0821	0.1791	0.0463	0.0803	0.1789
CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus)	0.0803	0.1789	0.0442	0.2094	0.0827
ME top [8]	0.2094	0.0827	0.0827	0.0725	0.2028
video CombMean (LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture)	0.0725	0.2028	0.0732	0.0753	0.1937
CombMax (LargSegm + ValSegm and Texture + MassVarSegm)	0.0753	0.1937	0.0619	0.0732	0.1937
CombMax (Edge + Texture and HSV + MassVarSegm)	0.0732	0.1937	0.0619	0.1845	0.0827
ME avg	0.1845	0.0827	0.0571	0.0723	0.1843
CombMax (Edge + Texture and HSV + MassVarSegm and HSL + Colorfulness)	0.0723	0.1843	0.0571	0.0737	0.1819
CombMax (LargSegm + ValSegm and Texture + MassVarSegm and Edge + Texture)	0.0737	0.1819	0.0564		

3.2.4 Conclusions In this chapter, we presented our participation [27] at the [MediaEval 2016 Predicting Media Interestingness Task](#) [48], that uses a set of traditional visual features and SVM-learning systems, a continuation of that work [30] on the same dataset that implements a set of aesthetic-based features and late fusion schemes, and the application of our aesthetic-based system [28] [on the MediaEval 2017 Predicting Media Interestingness Task](#) [47]. [To the best of our knowledge](#), our experimental results [with](#) 54 aesthetic [-based](#) systems [on](#) the 2016 image subtask still [represent the](#) current [state-of-the-art](#), therefore proving the value of such an approach and further exploring the correlations between visual aesthetics and interestingness. Furthermore, the improvements brought by the late fusion approaches can represent an important precedent for future developments.

### 3.3 Predicting violent scenes

#### 3.3.1 Introduction

[In this](#) section, [we present our](#) contribution to [the prediction of violent scenes in movies](#) and in YouTube 15 surveillance videos. This approach employs a ConvLSTM [182] structure that processes visual features created by processing video frame differences with a VGG [157] network. Experiments with this approach are

validated on two datasets: the MediaEval 2015 Violent Scene Detection dataset [158] and the VIF dataset [83].

### 3.3.2 Temporal deep learning systems

#### Motivation

The detection of violent scenes and events is an inherently temporal analysis; therefore, we choose to implement [state-of-the-art approaches with regards to the analysis of video sequences](#). While traditional methods based on motion features such as STIP and HMP [168] have been tested in literature, we wish to continue our work in analyzing top-performing systems [34], presented in Chapter 3.1.2, by adding this study that contains a state-of-the-art network.

#### Proposed approach

Our detection algorithm consists of an end-to-end temporal DNN with the ability to gather and recognize spatio-temporal information in video samples. The system does not directly use video frames as input for the processing stage, but differences between consecutive video frames, as proposed by [164]. By changing the input in this particular way, Sudhakaran et al. propose that the feature extracting networks will

15www.youtube.com 57

Figure 3.7: The diagram of the proposed solution. We highlight the main components, including the frame aggregator, VGG feature processor, ConvLSTM temporal aggregator and final FC layers. be trained from the start with an internal motion correlation between its hyperparameters. The frame differences are passed after the initial stage to a VGG-19 DNN model [157], which will encode a set of features for each pair of frame differences. In the final phase, ConvLSTM [182] layers will process the [output of the VGG network. The particular setup of the ConvLSTM layer](#) for this experiment is as follows. We use 256 filters with a dimension equal to  $3 \times 3$ , thus obtaining an output of 256 features for each processed video segment. Videos are processed with a variable-sized window of frames, equating to approximately 1 second. The final layers are [fully connected with a size of 512 neurons](#) each, and process the ConvLSTM output in order to obtain a final decision. This network architecture is presented in Figure 3.7.

#### Experimental setup

Experiments are carried out on two different datasets. The first one is the MediaEval 2015 Violent Scenes Detection task [158], which contains samples extracted from Hollywood-like movies. The composition of this dataset is detailed and described in Chapter 3.1.2. The second one is the VIF dataset [83], composed of short videos extracted from YouTube. In total, the VIF dataset is composed of 267 individual video files, with a total duration of 30 minutes, split into 246 files [in the training set, and 21 in the testing set](#). While this represents a much shorter dataset than MediaEval VSD, analyzing results in this setup will show how well the network can generalize to multiple data sources. Scenes in the VIF dataset are composed of crowd-based violence, being captured by normal security cameras, and the metric used by this dataset is Accuracy.

#### Experimental results

Experimental results are presented in Table 3.9, where they are also compared [with the current state-of-the-art performer on](#) each respective dataset. [The results for this approach are promising, with a maximum MAP value of 0.271 on the 2015 VSD dataset, representing a lower performance when compared with the current top result presented in \[37\], that achieves a MAP of 0.296, but with better results on the VIF dataset, i.e., an accuracy of 0.89, compared with the previous top results of 0.863 presented in \[67\].](#) Regarding the size of the window of frames, we tested values in {15, 20, 25, 30, 35}. While the best results, calculated on the VSD dataset are achieved for a window of size 30 (MAP = 0.271), a larger window of 35 frames also performs well, with a MAP of 0.270.

Table 3.9: Results of the proposed violence detection system, including comparison with [state-of-the-art results on the respective datasets](#). For [the MediaEval VD dataset results are](#) presented using the official MAP metric, while for the VIF dataset results are presented according to the Accuracy metric.

Method	Window config	VSD2015 (MAP)	VIF (Acc)
SOA	-	0.296 [37]	0.863 [67]
Our system	30	0.271	0.89

### 3.3.3 Conclusions

[In this chapter, we](#)

presented our approach for the task of predicting violent scenes in video samples. We developed an LSTM-based DNN, and tested the performance of this architecture on two datasets that target violence in Hollywood-like movies and surveillance videos extracted from YouTube. Results are promising, with the proposed approach performing above the current state of the art on the surveillance dataset.

### 3.4 Predicting media memorability

#### 3.4.1 Introduction

In this chapter, we present the contributions to the prediction of media memorability. Our paper [32] proposes the implementation of aesthetic and action recognition based systems to the memorability domain, and result augmentation via the implementation of a final late fusion step. My contributions to this work are represented by the implementation of the action recognition based systems and the implementation of late fusion schemes. Our approaches are tested on the publicly available dataset published during the MediaEval 2019 Predicting Media Memorability task [31].

#### 3.4.2 Action-based deep learning systems

**Motivation** In video processing, newly developed action recognition systems based on deep neural networks represent state-of-the-art approaches. These networks take advantage of temporal layers, such as LSTM layers [89], included in their architectures in order to produce better results on temporal data. Networks such as AssembleNet [139] or I3D [19] consistently represent state-of-the-art approaches at their moment of publication. We believe that using such networks would provide good results for the prediction of media memorability by accurately encoding temporal features associated with the video samples.

**Previous work** The concept of memorability has been intensely studied by researchers in psychology, computer vision, and human studies. The memorability of an image is shown to be an intrinsic propriety of the image [92, 103]. Furthermore, Shepard [151] shows that 61 human visual memory has a surprisingly massive storage capacity for memorizing visual samples. From a computer vision perspective, several methods for predicting the memorability of images and videos have been tested. Generally, high-level approaches are shown to have better performance [102] with regards to memorability prediction. Good results are also achieved with some DNN-based approaches, in [60] that use an AlexNet [107] based model for creating memorable video summaries, and [25] that use several models for frame-level memorability prediction or the MemNet approach of [102].

**Proposed approach** For this approach, we use several DNN models that are pre-trained on image aesthetics and action recognition. For the aesthetic based models, a ResNet-101 architecture [84] is fine-tuned on the memorability data. At the same time, for the action recognition DNNs the TSN [179] and I3D [19] networks are used as feature extractors and augmented with the C3D features [171] provided by the task organizers. Action recognition features are passed through a dimensionality reduction step, based on PCA, and training is processed via an SVR model. A final step involves the use of late fusion schemes. The outline of this approach is presented in Figure 3.8.

The aesthetic based architecture is described in Kang et al. [99], where the authors train the ResNet-101 architecture on the AVA dataset [120] for aesthetic value prediction. For the prediction of short and long-term memorability of videos, the model is fine-tuned on the memorability dataset, using key-frames extracted in two ways: (i) from the 4th, 5th and 6th second of each video and (ii) one frame from every second in the video. We extract the "Mixed 5" layer and use it as a feature from the I3D model, trained on the Kinetics dataset [100], while the "Inception 5" layer is extracted from the TSN model, trained on the UCF101 dataset [161].

We perform preliminary tests

Run	Model	Features
1	ResNet-101	I3D
2	I3D	I3D features
3	Fine-Tuning	TSN
4	Fine-Tuned	TSN features
5	Run	PCA SVR

Figure 3.8: The diagram of the proposed solution. We



represent the aesthetic-based network (ResNet-101) and action recognition networks (I3D, TSN, and the organizer provided C3D), their fine-tuning or extraction process and learning process, and the final late fusion (LF) stage. The components of the five individual runs submitted to the MediaEval Predicting Media Memorability task are also represented (Run 1 - Run 5). with regards to individual I3D and TSN features, but also with regards to their early fusion combinations with the provided C3D feature. These preliminary tests favor the early fusion combinations. Finally, an SVR model is used to train these features under a randomized 4-fold data split. We tune the parameters of this SVM model using an RBF kernel with C and gamma parameters taking values of 10k, where  $k \in [-4, \dots, 4]$ . Finally, we test three late fusion schemes that merge the action recognition systems' prediction outputs, as well as [the best action recognition system](#) with [the aesthetic-based](#) model prediction output. The three schemes are CombMax, CombMin and CombMean and they are implemented in the same way as presented in Chapter 3.2.3 Experimental setup Experiments are carried out on [the MediaEval 2019 Predicting Media Memorability task](#) [31]. [The setup of this dataset](#), including the number of samples and data splits, 63 Table 3.10: [Results of the proposed](#) memorability systems, including [preliminary tests on the devset and official results on the testset](#), according to the official Spearman's  $\rho$  metric. We also include a comparison with the top and average scores registered at the MediaEval task. The five runs are denoted r1 - r5. Run System description Devset - Spearman's  $\rho$  short-term long-term [Testset - Spearman's  \$\rho\$  short-term long-term](#) r5 r2 r4 r1 r3 ME top [7, 134] LF Aesthetic + Action (r1 + r2) Action (TSN + I3D) ME avg LF Action (r2 + r3) Aesthetic Action (C3D + I3D) - 0.494 0.473 - 0.466 0.448 0.433 - 0.265 0.259 - 0.200 0.230 0.204 0.528 0.477 0.450 0.448 0.439 0.401 0.386 0.277 0.232 0.228 0.206 0.218 0.203 0.184 is presented and detailed in Chapter 3.1.3. A full comparison with results from other participants to the MediaEval benchmarking competition can also be found in that section of the thesis. Experimental results The experiments are again carried out in two stages. While the first stage represents the development and validation of the systems on the devset, the second stage represents the deployment of the selected systems on the testset. Results are presented in Table 3.10, where they are also compared with the top performer and the average scores from the MediaEval task. As previously mentioned, several variations are used [in the training stage of the aesthetic DNN approach](#). [For the short-term memorability task](#), two training approaches, i.e., training with the 5th frame and training with one frame per second, produce similar scores, with a Spearman  $\rho = 0.448$ , while in the long-term memorability task using the 5th frame produces better results, with a Spearman  $\rho = 0.230$ . Considering that the large majority of videos in this dataset present only one visual scene, a more extensive training dataset, as is the case for the multi-frame approach, may not be beneficial, as all frames in a video could be similar. In general, however, there is little difference between the results reported for these different approaches. 64 For the action recognition systems, individual systems are outperformed by early fusion schemes. Results for individual systems on the devset are as follows for the short-term memorability: TSN  $\rho = 0.418$ , I3D  $\rho = 0.401$  and C3D  $\rho = 0.3521$ . This performance further drops when the PCA processing is not implemented, therefore proving the positive influence of dimensionality reduction schemes. The top 2 performing early fusion combinations on the devset of action recognition network features are as follows: TSN + I3D, with  $\rho = 0.473$  [for short-term](#) and  $\rho = 0.259$  [for long-term](#) and C3D + I3D, with  $\rho = 0.433$  on short-term and  $\rho = 0.204$  on long-term. We propose two late fusion combinations, namely one that would merge the two best action recognition approaches and another one that would merge both the aesthetics and the action

recognition approaches. As shown in previous experiments presented in this work, CombMean and CombMax produce better results than their inducers, with CombMean being the best performing late fusion scheme. The final results on the testset, shown in Table 3.10, show that the best performing system uses a late fusion combination of aesthetic network prediction outputs and action recognition early fusion prediction outputs. Two of our runs perform above the MediaEval average results, namely the early fusion of action features represented by the TSN and I3D and the late fusion approach that merges action and aesthetic results. For the latter, the best results are  $\rho = 0.477$  [for short-term memorability](#) and  $\rho = 0.232$  [for long-term memorability](#).

### 3.4.3 Conclusions

In this chapter, we presented our participation at the MediaEval 2019 Predicting Media Memorability task [31], that uses aesthetics and action recognition based networks for predicting short and [long term memorability for video samples](#). The results recorded during this competition are promising and continue to enforce the idea that late fusion systems can be successfully applied in order to increase the results of their individual inducers significantly.

### 3.5 Late fusion with deep ensemble systems

#### 3.5.1 Introduction

In this chapter, we present the contributions to the creation of deep ensembling systems. Our works [162] and [29] propose the creation of ensemble systems that use DNNs as the main ensembling driver. [To the best of our knowledge, this type of approach](#) represents a novelty in the field of information fusion, where so far, DNNs have only been used as inducers for traditional fusion systems. My contribution to this work is represented by (i) the creation of two novel 2-D and 3-D input transformation schemes that allow the use of multidimensional deep neural layers, (ii) the implementation of convolutional layers in ensembling systems, (iii) and the creation of a novel DNN layer, specially designed for fusion systems, called the Cross-Space-Fusion layer. The proposed systems are tested on several publicly available datasets published as part of several MediaEval tasks, using as inducers the systems that participated at their respective tasks, as provided to us by the task organizers.

#### 3.5.2 Motivation

As presented in some of the previous chapters, ensembling or late fusion systems seem to be able to significantly increase the performance of inducer algorithms for subjective tasks such as visual interestingness 3.2.3 and memorability 3.4.2 prediction. Our findings in this domain are supported by other works, where ensembles managed [to achieve state-of-the-art results](#). Examples regarding [this](#) would include video interestingness [180], video memorability [7], and emotional content analysis [165], but also domains that do not deal with such subjective concepts, examples here including the classification of human actions in videos [163]. While these approaches do use several late fusion schemes, they do so on a lower scale, using few inducers or testing 16Paper under major review 67 their approaches on a single dataset. Furthermore, the ensembling schemes proposed by these authors are mostly represented by statistical methods, and we believe that using more inducers and employing better, more modern, ensembling schemes will significantly improve performance.

#### 3.5.3 Previous work

So far, ensembling systems have employed a set of traditional methods for driving the ensemble. Some examples are already presented in this thesis, mainly statistical methods such as CombMin, CombMax, CombMean, etc. Other popular methods from the literature include boosting methods such as AdaBoost [64] and Gradient Boosting [65], bagging methods [14], methods based on random forests [15]. For a more comprehensive understanding of late fusion systems, we refer the reader to some literature survey papers that deal with this subject [70, 106, 140]. Many taxonomies of ensembles have been proposed, taking into account the main ensembling method or inducer properties like combination [52, 148, 108], inducer diversity [16], inducer dependency [140], and size of the ensemble [137]. However, as we previously

mentioned, our approach would represent a novelty with regards to the introduction of DNN algorithms as the primary ensembling method and with regards to the number of systems employed by the ensemble.

### 3.5.4 Proposed approach

For any standard ensemble, given a set  $S$  of  $M$  samples  $s_i, i \in [1, M]$  and a set  $F$  of  $N$  classifier or regression inducer algorithms  $f_i, i \in [1, N]$ , each algorithm will produce an output for every given sample  $y_{i,j}, j \in [1, N], i \in [1, M]$ , as follows:

$$Y = \begin{bmatrix} y_{1,1} & \dots & y_{1,N} \\ \vdots & \ddots & \vdots \\ y_{M,1} & \dots & y_{M,N} \end{bmatrix} \quad S = s_1 \ s_2 \ \dots \ s_M \quad F = f_1 \ f_2 \ \dots \ f_N \quad (3.2)$$

Ensembling involves the creation of an algorithm  $E$ , that aggregates the outputs of inducers and learns patterns in a training set composed of individual inducer outputs and ground truth data. These patterns are then applied on the validation set, in order to produce a new output for new samples,  $e_i, i \in [1, \dots, M]$ , that represents a better-tuned output with regards to metric. The value space of  $e_i$  can differ according to the type of task the ensemble attempts to solve. For example, in regression tasks  $e_i \in [0, 1]$  or  $[-1, 1]$ , while in binary classification tasks those values can be 0 or 1. Furthermore, for multi-label or multi-class classification,  $e_i$  will be represented by a vector of values, of equal size to the number of possible classes or labels. The proposed DeepFusion approach deploys several types of DNN that take as input the set of inducer outputs  $Y$  and produces a new set of ensembled outputs  $e$ , according to the positive and negative biases the DNN managed to learn during the training process. We thus propose to start with the creation of a baseline deep ensembling system, based on a combination of variable-sized dense layers. This baseline will then be augmented by the addition of convolutional layers, and finally, with the addition of the novel Cross-Space-Fusion (CSF) layer. While dense based networks use a 1-dimensional input for each image and video sample, convolutional and CSF layers use 2-dimensional or 3-dimensional inputs. The purpose of these layers is similar to the purpose of convolutions in image processing: we will attempt to discover and learn spatial correlations and patterns between input values that are spatially grouped together. However, such information is impossible to extract from a 1-D vector of inputs that corresponds to each sample, created by the outputs of individual inducers. We, therefore, create a set of input transformation schemes that allow us to build 2D and 3D input structures, based on the similarity degree between individual inducers, thus making possible the implementation of convolutional and CSF layers.

**Dense networks** Dense layers are known for being able to classify input data into output categories accurately, thus representing an integral part of all DNN approaches. Considering their input-agnostic nature, we theorize that building an initial baseline network that integrates several dense layers would represent a valuable starting point in creating the network. A representation of a dense ensembling architecture is presented in Figure 3.9. We choose to vary a set of parameters of these networks in order to optimize its performance with regards to the tasks being studied. The following parameters are chosen: (i) number of layers, with values of  $\{5, 10, 15, 20, 25\}$ ; (ii) the number of neurons per layer, with values of  $\{25, 50, 500, 1000, 2000, 5000\}$ ; and (iii) the presence or absence of batch normalization layers. We change the values of these parameters until the best results on the chosen datasets are achieved. ... ..

... Output Input Dense B.N. Figure 3.9: DeepFusion dense network architecture (DF-Dense): variable number of layers, number of neurons per layer and the presence or absence of Batch Normalization (BN) layers. Input decoration We choose to pre-process the input data and decorate each element with output scores and data from the most similar inducers to generate spatial information. Given an image or video sample  $s_i, i \in [1, M]$ , each of the  $N$  inducer algorithms will produce a set of scores,  $Y_i$ , as described in Equation 3.3, and, as mentioned before, this kind of input has no intrinsic spatial correlation associated

with it. In the first step of the input pre-processing technique, we analyze the correlation between the individual inducers  $f_i$ ,  $i \in [1, N]$ . This correlation can be determined by any standard method, such as Pearson's correlation score. However, to ensure an optimized learning process, we will use the same metric as the one the task uses as its official metric. Given any  $f_i$ ,  $i \in [1, N]$  inducer system, that produces the vector  $f_i^-$  of outputs across the entire set of samples, as described in Equation 3.4, and a vector of correlation scores  $cri$  between this inducer and all the other inducers can be generated as presented in Equation 3.5. To generate an appropriate spatial correlation, we must use the descending ordered version of this vector, denoted  $crd_i$ :  $Y_i = [y_{1,i} \ y_{2,i} \ \dots \ y_{N,i}]$ ,  $f_i^- = [f_{i,1}^- \ f_{i,2}^- \ \dots \ f_{i,M}^-]$  (3.4)  $cri = [cr_{1,i} \ cr_{2,i} \ \dots \ cr_{N-1,i}]$  (3.5) As we previously mentioned, we consider both a 2D and 3D representation of the decorated input space. For the 2D representation, named  $tr2D$ , we apply Equation 3.6, and this input decoration scheme will be used for decorating the input for convolutional network usage. On the other hand, the two Equations presented in 3.7 describe the 3D representation,  $tr3D$ , with each of the two matrices being stored at different indexes in the 3rd dimension, creating a structure used by the CSF layer.  $tr2D_{i,j} = [r_{1,i,j} \ r_{2,i,j} \ \dots \ r_{N-1,i,j}]$ , (3.6)  $tr3D_{i,j} = [r_{1,i,j} \ r_{2,i,j} \ \dots \ r_{N-1,i,j}]$  (3.7) In this example, each element  $si,j$ , representing the prediction output produced by inducer  $i$  for a sample  $j$  of the input to our neural network model, is decorated with scores from similar systems,  $c_{1,i,j}$  representing the most similar system,  $c_{2,i,j}$  representing the second most similar system and so on. For the  $r$  values of our new matrix we input the correlation scores for the most similar system ( $r_{1,i,j}$ ), the second most similar ( $r_{2,i,j}$ ) and so on, with the value 1 as centroid, corresponding to the initial  $si,j$  element. The outline of the 3D decoration method is presented in Algorithm 1. The spatial dimension per media sample, before the decoration process is  $N$ , in other words, equal to the number of inducer systems deployed. For the 2D approach, this dimension grows to  $(3 \times N, 3)$ , while for the 3D approach the size is  $(3 \times N, 3, 2)$ . Dense networks with convolutional layers [A general presentation of the employed convolutional architecture is depicted in Figure 3.10](#). After processing the input and transforming it into a  $tr2D$  form, this input is fed into a convolutional layer. Given the  $3 \times 3$  padding of each element of the original input, we also choose to use a  $3 \times 3$  filter in our proposed architecture, therefore obtaining 10 trainable parameters in this layer. We use a stride parameter of 3, ensuring that each convolutional filter only processes similar systems. This structure [is followed by an average pooling layer](#) that will bring the output of the convolution to 72 [Algorithm 1:](#)

```

Input pre-processing algorithm for inducer i, sample j
Data:  $i, j, si,j, Y_i = [y_{1,i} \ y_{2,i} \ \dots \ y_{N,i}]$ ,  $f_i^- = [f_{i,1}^- \ f_{i,2}^- \ \dots \ f_{i,M}^-]$ 
Result:  $C_{i,j}, R_{i,j}$ 
begin [ ] [ ] //create the empty structures;  $tr3D_{c,i,j}, tr3D_{r,i,j} \leftarrow \text{zeros}(3, 3)$ ;
//compute the crm correlations; for  $m \leftarrow 0$  to  $M$  do  $crm \leftarrow \text{zeros}(M - 1, 2)$ ; //compute the crm
correlations for each inducer; if  $m! = i$  then  $crm[m, 0] \leftarrow \text{CalcCorrelation}(f_i^-, f_m^-)$ ;  $crm[m, 1] \leftarrow m$ ; end
end //order the inducers according to their correlation;  $crm \leftarrow \text{Sort}(crm)$ ; //append the values to the 2-D
structures according to Eq. 3.7; for  $k \leftarrow 1$  to 8 do  $tr3D_{c,i,j} \leftarrow \text{AppendStructure}(Y_i[crm[k, 1]], k)$ ;  $tr3D_{r,i,j} \leftarrow \text{AppendStructure}(crm[k, 0], k)$ ; end //append the central values;  $tr3D_{c,i,j} \leftarrow \text{AppendStructure}(si,j, 0)$ ;
 $tr3D_{r,i,j} \leftarrow \text{AppendStructure}(1, 0)$ ; return  $tr3D_{c,i,j}, tr3D_{r,i,j}$ ; end
the initial 1D input shape. We also test 1, 5, and 10 filters per convolution, allowing the network to perform a more extensive analysis of the similarities. Dense networks with Cross-Space-Fusion layers Finally, we introduce the Cross-Space-Fusion (CSF) layer, whose general design is presented in Figure 3.11. This layer takes the 3D  $tr3D$  array and, for

```





classification task, consisting of 44 training/validation movies, with a total duration of more than 15 hours, and 12 testing movies with a total duration of approx. 9 hours. On the other hand, the ImageCLEFmed 2019 Concept Detection is an automatic multi-label classification image captioning and scene understanding data set [126] consisting of 56,629 training, 14,157 validation, and 10,000 test radiology examples with multiple classes (medical concepts) associated with each image. In total, there are 5,528 unique concept identifiers, whereas the distribution limits per images [in the training, validation, and test sets are](#) between 1-72, 1-77, and 1-34 concepts, respectively. In order to create an adequate baseline of inducers, we used systems submitted to the respective tasks as inducer systems, as they have been provided to us by the task organizers. Other setups would be impractical, as they would involve training a large number of systems from the start, and considering that many times the authors of the proposed systems do not provide their source code. Furthermore, task organizers are only able to provide us the system runs from the testset. Therefore we decided to create two types of splits on this inducer output data. In this regard, the split samples are randomized, and 100 partitions are generated to assure thorough coverage of the data, using two protocols: (i) 75% training and 25% testing (KF75), and (ii) 50% training and 50% testing (KF50). In order to avoid random splits that favor our type of approach, we generate 100 of these partitions and report the results as average values between the 100 runs. We would like to point out that, while this does not represent an accurate duplication of the original dev/test split, it does represent a disadvantage for our training stage, as we will train our deep learning fusion methods on less data than the original systems submitted to the MediaEval and ImageCLEF tasks. In this respect, we used the following number of inducers for each of the experimental datasets: - INT2017.Image - 33 systems, - INT2017.Video - 42 systems, - VSD2015.Video - 48 systems, - Aro2018.Video and Val2018.Video - 30 systems, - Fear2018.Video - 18 systems, - Capt2019.Image - 58 systems.

3.5.6 Experimental results [In the following section, we will present the results of our experiments.](#) For each task and set of experiments, we will provide two baselines. The first one is composed of the top-performing systems, for each dataset, recorded both during the corresponding MediaEval competition and outside of it, in [state-of-the-art](#) works. [The second set of baseline experiments](#) are represented by a set of traditional ensembling systems that include: CombMax, CombMean, CombMean, CombAvg, CombSum, presented in Table 3.11: Final results for the convolutional architecture (DF-Conv) experiments. [These results are compared with the best results from the MediaEval competition \(ME top\), best results from the state-of-the-art literature \(SOA top\), the best results from the baseline fusion systems \(Emb top\) and with the best results of the dense architecture experiments \(DF-Dense\) for the INT2017.Image task \(with official MAP@10 metric\), INT2017.Video \(with official MAP@10 metric\) and VSD2015.Video \(with official MAP metric\). The dataset split \(dev/test, KF50 or KF75\) used to produce the results is also presented.](#)

Dataset	ME top	SOA top	Emb top	DF-Dense	DF-Conv
INT2017.Image (MAP@10)	0.1385 [131]	0.156 [125]	0.1523	0.1674	0.2316
INT2017.Video (MAP@10)	0.0827 [8]	0.093 [180]	0.0961	0.1129	0.1563
VSD2015.Video (MAP)	0.296 [37]	0.303 [113]	0.3521	0.392	0.6192

0.6341 0.6281 0.6471 in previous chapters, and two boosting strategies, namely AdaBoost [64] and gradient boosting [65]. Results for the convolutional architecture For the convolutional architectures, we run tests on the INT2017.Image, INT2017.Video and VSD2015.Video [datasets. The results are presented in Table 3.11.](#) While the traditional early fusion schemes did improve the results,

with AdaBoost being the best performer for the INT2017.Image and INT2017.Video datasets and Gradient boosting being the best performer for the VSD2015.Video dataset, their improvements are still small, especially for the KF50 setup. On the other hand, both deep ensembling architectures significantly increase performance. The best performer in these tests is the DF-Conv architecture. As we mentioned in the description of the training protocol, the DF-Conv is built upon the best performing DF-Dense architecture, in order to analyse if the addition of convolutional layers over an already saturated dense architecture can make a difference with regards to results. Thus, the best performing DF-Dense architectures are as follows: (i) for INT2017.Image the best DF-Dense system uses 10 dense layers with 1000 neurons per layer and no BN integration, attaining MAP@10 values of 0.2316 for KF50 and 0.3355 for KF75; (ii) for INT2017.Video the best DF-Dense system has 25 78 layers with 2000 neurons each and BN layers, with MAP@10 performance of 0.1563 for KF50 and 0.2677 for KF75; (iii) finally, for VSD2015.Video, best performance is achieved with 5 dense layers with 500 neurons each and no BN layers, achieving a MAP score of 0.6192 for KF50 and 0.6341 for KF75. Finally, with a single exception, namely the INT2017.Image KF50 configuration, all the DF-Conv architectures improved the results of their DF-Dense counterparts. The best performing DF-Conv architectures used 5 filters for INT2017.Image and INT2017.Video and 10 filters for VSD2015.Video. The best results for these datasets are as follows: (i) for INT2017.Image MAP@10 values of 0.2293 in a KF50 configuration and 0.3436 for KF75, (ii) for INT2017.Video MAP@10 values of 0.1692 for KF50 and 0.2799 for KF75, (iii) and finally, for VSD2015.Video, MAP values of 0.6281 and 0.6471 in KF50 and KF75 configurations respectively. These results represent a significant increase in performance, both over the ME top systems, that also represented inducers for our system, but also over state-of-the-art results, namely 120% for INT2017.Image, 200.9% for INT2017.Video and 113.5% for VSD2015.Video. Results for the Cross-Space-Fusion architecture For the CSF architecture we run tests on the Aro2018.Video, Val2018.Video, Fear2018.Video and Capt2019.Image. The results of these experiments are presented in Table 3.12. Considering that these particular datasets and tasks are newer than the ones selected for DF-Conv architecture, no state-of-the-art systems have yet been developed for them, therefore we cannot use SOA top as a comparison baseline. Just like the case for the DF-Conv architectures, the improvements brought by the traditional fusion systems are minimal. Gradient boosting provides the best results for Val2018.Video, while AdaBoost achieves best performance for the rest of the datasets.

79 Table 3.12: Final results for the CSF architecture (DF-CSF) experiments. [These results are compared with the best results from the](#) MediaEval competition (ME top), the best results from the baseline fusion systems (Emb top) and with the best results of the dense architecture experiments (DF-Dense) for the Aro2018.Video task (with official MSE and PCC metrics), Val2018.Video (with official MSE and PCC metrics), Fear2018.Video (with official IoU metric) and Capt2019.Image (with official F1 metric). The dataset split (dev/test, KF50 or KF75) used to produce the results is also presented. Dataset ME top Emb top DF-Dense DF-CSF Dataset split dev/test

Dataset	ME top	Emb top	DF-Dense	DF-CSF	Dataset split	dev	test
Aro2018.Video (MSE)	KF50	0.1334	0.1321	0.1253	0.0571	0.0549	0.0568
	KF75	0.0543	0.0543	0.3358	0.3547	0.3828	0.8018
Aro2018.Video (PCC)	KF50	0.8315	0.8073	0.8422	0.8073	0.8422	0.8422
	KF75	0.0837	0.0814	0.0769	0.0640	0.0626	0.0636
Val2018.Video (MSE)	KF50	0.0625	0.0625	0.3047	0.3372	0.3972	0.7876
	KF75	0.8101	0.7903	0.8123	0.1575	0.1597	0.1733
Val2018.Video (PCC)	KF50	0.1938	0.2129	0.2091	0.2242	0.2242	0.2242
	KF75	0.2823	0.2804	0.2846	0.3462	0.3740	0.3610
Fear2018.Video (IoU)	KF50	0.3912	0.3912	0.3912	0.3912	0.3912	0.3912
	KF75	0.2823	0.2804	0.2846	0.3462	0.3740	0.3610
Capt2019.Image (F1)	KF50	0.2823	0.2804	0.2846	0.3462	0.3740	0.3610
	KF75	0.2823	0.2804	0.2846	0.3462	0.3740	0.3610

Again both deep ensemble architectures significantly outperform other results. The best performing DF-Dense architectures are the

as follows: (i) for both the arousal and valence datasets, we use a 5 layer architecture with 500 neurons per layer and BN layers; (ii) for Fear2018.Video the best performing architecture employs 10 dense layers with 500 neurons, without BN integration; (iii) finally, for Capt2019.Image again the best performing architecture uses 5 layers with 500 neurons and no BN. Regarding the CSF architecture, the results are further improved when compared with the DF-Dense approach. For the arousal and valence runs, the optimal tr3D setup is 4S, with only 4 similar systems used for decorating the input. MSE results are improved by 59.3% and 25.3% for the KF75 setup, with regards to MSE. However, the starting ME top results are already quite high, therefore a huge improvement, like the ones shown in the previous section are impossible. On the other hand, when looking at the PCC metric, the improvements are much larger, with 150.1% and 166.6%. For the Fear2018.Video, the DF-CSF architecture improves results by 42.3%, while for the Capt2019.Image data, improvements are at 38.6%.

### 3.5.7 Conclusions

In this chapter, we discussed the creation of a deep ensemble framework, that represents a novel research direction with regards to late fusion approaches. Our systems use dense and convolutional layers for combining inducer predictions, as well as the novel Cross-Space-Fusion layer. We also introduced two novel input transformation schemes that allow the implementation of convolutional and CSF architectures on inducer predictions. Our systems are tested on seven datasets that cover several types of machine learning tasks, including regression, binary classification, and multi-label classification, and provided significant improvements both over current state-of-the-art approaches and over traditional late fusion systems.

## Chapter 4 General conclusions and perspectives

### 4.1 Contributions and publications

In this chapter I will summarize the main personal contributions to research papers published during my doctoral research program. These contributions are as follows:

- In (P2) I proposed the implementation of a set of traditional visual features for the prediction of media [interestingness](#). [Experimental validation is performed on the MediaEval 2016 Predicting Media Interestingness dataset](#).
- In (P3) and (P6) I proposed the implementation of a large set of finely-grained aesthetic features, based on color, texture, photographic and composition rules, [for the prediction of media interestingness](#). The methods [are validated both on the 2016 and on the 2017 versions of the MediaEval Predicting Media Interestingness datasets](#), as well as the implementation of early and late fusion schemes for performance optimization. To the best of my knowledge, the results recorded on the 2016 image subtask still represent the state-of-the-art with regards to MAP performance.
- In (P8), (P10), (P11), (P12) I proposed the implementation of visual methods for the creation of movie recommending systems. These research papers also produced the MMTF-14K dataset, where I provided a set of aesthetic and DNN-based descriptors as baselines for researchers that wish to use our dataset.
- (P13) currently represents the largest literature review on the prediction of media interestingness and its covariates. My contributions to this work are related to the study of computer vision approaches to the prediction of interestingness and its correlated concepts, the creation of a taxonomy model that studies the positive, negative and still unexplored correlations between interestingness and other subjective concepts, and, with a lower degree of involvement, the study of human understanding of interestingness.
- In (P14) I was the main organizer of the MediaEval 2019 Predicting Media Memorability task.
- In (P15) I proposed the implementation of [action recognition based DNNs for the prediction of media memorability](#). Results are validated on the MediaEval 2019 Predicting Media Interestingness, and early and late fusion schemes are deployed for performance optimization.
- (P18) represents a thorough analysis of the VSD96 dataset, aimed at the detection of

violent video scenes. My main contributions to this work are represented by the overall analysis of the methods employed on this dataset by a large number of authors, a study of the influence of features on the prediction results and formulating some of the main conclusions with regards to the prediction of violence. - (P19), a work currently under review, represents a thorough analysis of the Interestingness10k dataset, aimed at the prediction of image and video interestingness. My [main contributions to this paper are](#) as follows: [the analysis of the overall performance of systems that use this dataset](#), an [analysis of the influence of features on the performance of systems](#), [the study of the generalization capabilities of systems and recommendations with regards to system performance](#). Some shared contributions include: the study of state-of-the-art DNN approaches and interpretability of results, as well as the deployment of statistical, boosting and DNN-based late fusion systems for the improvement of the results recorded during the MediaEval 2016 and 2017 editions of the Predicting Media Interestingness task. - (P17) represents a novel approach with regards to ensembling systems. The novelty here is represented by the introduction of DNN architectures as the main ensembling method for combining inducer prediction output. My main contributions to this paper are represented by the creation of an input decoration method, that facilitates a spatial grouping of similar inducers and by the implementation of convolutional layers for processing the decorated input. Validation is carried out on three regression tasks, namely the MediaEval 2017 image and video subtasks from [the Predicting Media Interestingness task](#), and [the 2015 MediaEval Violent Scenes Detection task](#), and, as results show, these methods greatly improve system performance. This work has continued, but newer results are currently unpublished. Newer results include the addition of another novel input decoration model, as well as the introduction of a novel DNN layer called Cross-Space-Fusion that is specially designed for processing ensemble data. (P1) 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. (P2) [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. 85 (P3) [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. (P4) [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. (P5) 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. (P6) [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. (P7) 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. (P8) [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. (P9) 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. 86 (P10) [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. (P11) [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. (P12) [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: 3.4). (P13) [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: 6.131). (P14) [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. (P15) [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. 87 (P16) [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](https://doi.org/10.1007/978-3-030-45442-5_69), ECIR 2020 Proceedings, April 14-17, Lisbon, Portugal, 2020. (P17) [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. (P18) [M.G. Constantin, L.D. Ștefan, 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: 6.28). (P19) Paper under major review : [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. 4.2 Conclusions [This thesis presents](#) my personal [contributions to the](#) automatic analysis [of the](#) visual impact of multimedia data, with an accent on the study of interestingness, aesthetics, memorability, violence and affective value and emotions. Chapter 2 presents an analysis of the [current state-of-the-art with](#) regards [to](#) concept taxonomy [and](#) definitions, theories on the human understanding of subjective multimedia properties, datasets and user studies, computational approaches, and current applications and future perspectives on the use of these properties. Chapter 3 presents my contributions to this field. The first part of this chapter covers the datasets and benchmarking initiatives I have contributed to. Following this, the thesis presents several computer vision methods developed



during my doctoral program and analyses the contributions to the current computational landscape brought by these methods. Methods presented here are related to: (i) the prediction of media interestingness via traditional visual features in an SVM learning setting, and the implementation of aesthetic-based features and statistical late fusion schemes for interestingness prediction; (ii) the detection of violent scenes via the implementation of a ConvLSTM approach; (iii) the prediction of media memorability with the help of action recognition deep neural networks; (iv) the creation of a novel deep learning based approach to ensemble learning, the creation of new input decoration methods that would allow the processing of correlated inducers in our deep fusion systems and a novel type of deep neural network layer, the Cross-Fusion-Layer, specially designed for the processing of ensemble systems. The results presented in this thesis are promising, especially considering that the proposed deep fusion systems significantly increase state-of-the-art performance. While in general late fusion systems do require more processing power, given that the data is processed by multiple inducers, one must consider that these types of approaches will prove to be useful, given the constant improvement of GPU processing power and the advent of online services dedicated to processing massive amounts of 90 data in a reasonable time. I consider that such systems can be deployed in many use cases, mainly in scenarios where the results of individual systems are not good enough for a final market-ready solution, or in critical infrastructure systems, where accurate results are more important than the cost of a system.

#### 4.3 Future perspectives

In continuation of this work, the most important aspect would be the implementation of systems that are better tuned for their respective tasks. Some examples are already presented in this thesis, i.e., aesthetic-based features, but I consider that, by implementing more of these types of systems based on previous research from the fields of psychology and behaviour analysis, better architectures can be constructed and their results would better benefit the multimedia community. Furthermore, given the results of the deep ensemble system, I consider that it represents a very interesting research direction for the future. While this approach represents, [to the best of my knowledge, the first attempt](#) in creating such deep fusion systems, future developments may include: the creation of novel input decoration methods, the addition of novel layers and training schemes for the existing layers, and studies with regards to optimizing the collection of employed inducers.

Bibliography [1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos. In IEEE International Workshop on Content-Based Multimedia Indexing, pages 1–6. IEEE, 2015. [2] Peter P Aitken. Judgments of pleasingness and interestingness as functions of visual complexity. *Journal of Experimental Psychology*, 103(2):240–244, 1974. [3] Jurandy Almeida. Unifesp at mediaeval 2016: Predicting media interestingness task. In Working Notes Proceedings of the MediaEval 2016 Workshop., 2016. [4] Richard Chase Anderson. Interestingness of children’s reading material. Center for the Study of Reading Technical Report; no. 323, 1984. [5] Hannah Arendt. *On violence*. Houghton Mifflin Harcourt, 1970. [6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. [7] David Azcona, Enric Moreu, Feiyan Hu, Tomás Ward, and Alan Smeaton. Predicting media memorability using ensemble models. In Working Notes Proceedings of the MediaEval 2019., 2019. [8] Oifa Ben-Ahmed, Jonas Wacker, Alessandro Gaballo, and Benoit Huet. Eu-recom@ mediaeval 2017: Media genre inference for predicting media interest- ingness. In Working Notes Proceedings of the MediaEval 2017 Workshop., 2017. [9] Daniel E Berlyne. ‘interest’ as a psychological concept. *British*

journal of psychology. General section, 39(4):184–195, 1949. [10] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960. [11] Daniel E Berlyne. Novelty, complexity, and hedonic value. *Perception & Psychophysics*, 8(5):279–286, 1970. [12] Daniel E Berlyne. *Aesthetics and psychobiology*, volume 336. JSTOR, 1971. 93 [13] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008. [14] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. [15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. [16] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. [17] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision re- search*, 116:165–178, 2015. [18] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002. [19] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. [20] Christel Chamaret, Claire-Helene Demarty, Vincent Demoulin, and Gwenaëlle Marquant. Experiencing the interestingness concept within and between pictures. *Electronic Imaging*, 2016(16):1–12, 2016. [21] Ang Chen, Paul W Darst, and Robert P Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3):383–400, 2001. [22] Chih-Ming Chen and Ying-Chun Sun. Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. *Computers & Education*, 59(4):1273–1285, 2012. [23] Sharon Lynn Chu, Elena Fedorovskaya, Francis Quek, and Jeffrey Snyder. The effect of familiarity on perceived interestingness of images. In *IS&T/SPIE Electronic Imaging*, volume 8651, pages 86511C–86511C. International Society for Optics and Photonics, 2013. [24] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability task. *Working Notes Proceedings of the MediaEval 2018 Workshop.*, 2018. [25] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. Transfer learning for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*, 2018. [26] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engelberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2531–2540, 2019. 94 [27] Mihai Gabriel Constantin, Bogdan Boteanu, and Bogdan Ionescu. Lapi at mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016. [28] Mihai Gabriel Constantin, Bogdan Andrei Boteanu, and Bogdan Ionescu. Lapi at mediaeval 2017-predicting media interestingness. In *Working Notes Proceedings of the MediaEval 2017 Workshop.*, 2017. [29] Mihai Gabriel Constantin, Liviu-Daniel Ștefan, Bogdan Ionescu, Ngoc Q. K. Duong, Claire-Hélène Demarty, and Mats Sjöberg. Visual interestingness prediction: A benchmark framework and literature review. Under Review at: *International Journal of Computer Vision*, 2020. [30] Mihai Gabriel Constantin and Bogdan Ionescu. Content description for predicting image interestingness. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017. [31] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc QK Duong, Xavier Alameda-Pineda, and Mats Sjöberg. Predicting media memorability task at mediaeval 2019. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019. [32] Mihai Gabriel Constantin, Chen Kang, Gabriela Dinu, Frédéric

Dufaux, Giuseppe Valenzise, and Bogdan Ionescu. Using aesthetics and action recognition-based networks for the prediction of media memorability. In Working Notes Proceedings of the MediaEval 2019 Workshop., 2019. [33] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)*, 52(2):1–37, 2019. [34] Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Hélène Demarty, Mats Sjöberg, Markus Schedl, and Guillaume Gravier. Affect in multi-media: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, 2020. [35] David Culbert. Television’s visual impact on decision-making in the usa, 1968: The tet offensive and chicago’s democratic national convention. *Journal of Contemporary History*, 33(3):419–449, 1998. [36] Qi Dai, Zuxuan Wu, Yu-Gang Jiang, Xiangyang Xue, and Jinhui Tang. Fudan-just at mediaeval 2014: Violent scenes detection using deep neural networks. In Working Notes Proceedings of the MediaEval 2014 Workshop., 2014. [37] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In Working Notes Proceedings of the MediaEval 2015., 2015. 95 [38] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006. [39] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *IEEE International Conference on Image Processing*, pages 105–108. IEEE, 2008. [40] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 455–459, 2018. [41] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. Mmtf-14k: a multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 450–455, 2018. [42] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. Using visual features based on mpeg-7 and deep learning for movie recommendation. *International journal of multimedia information retrieval*, 7(4):207–219, 2018. [43] Emmanuel Dellandréa, Martijn Huigslot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, and Mats Sjöberg. The mediaeval 2018 emotional impact of movies task. In Working Notes Proceedings of the MediaEval 2018 Workshop., 2018. [44] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The mediaeval 2011 affect task: Violent scene detection in hollywood movies. In Working Notes Proceedings of the MediaEval 2011 Workshop., 2011. [45] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The mediaeval 2012 affect task: Violent scene detection. In Working Notes Proceedings of the MediaEval 2012 Workshop., 2012. [46] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang, and Yu-Gang Jiang. The mediaeval 2013 affect task: violent scenes detection. In Working Notes Proceedings of the MediaEval 2013 Workshop., 2013. [47] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Duong. Mediaeval 2017 predicting media interestingness task. In Working Notes Proceedings of the MediaEval 2017 Workshop., 2017. [48] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc QK Duong, and Frédéric Lefebvre. Mediaeval 2016 predicting media interestingness task. In Working Notes Proceedings of the MediaEval 2016 Workshop., 2016. 96 [49] Nadia Derbas, Bahjat Safadi, and Georges Quénot. LIG at

mediaeval 2013 affect task: Use of a generic method and joint audio-visual words. In Proceedings of the MediaEval 2013 Workshop, 2013. [50] Arturo Deza and Devi Parikh. Understanding image virality. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1818–1826. IEEE, 2015. [51] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1657–1664. IEEE, 2011. [52] Robert PW Duin. The combining classifier: to train or not to train? In Object recognition supported by user interaction for service robots, volume 2, pages 765–770. IEEE, 2002. [53] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. [54] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008. [55] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595, 2003. [56] Goksu Erdogan, Aykut Erdem, and Erkut Erdem. HUCVL at mediaeval 2016: Predicting interesting key frames with deep models. In Working Notes Proceedings of the MediaEval 2016 Workshop., 2016. [57] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6363–6372, 2018. [58] Shaojing Fan, Tian-Tsong Ng, Bryan L Koenig, Ming Jiang, and Qi Zhao. A paradigm for building generalized models of human image perception through data fusion. In IEEE Conference on Computer Vision and Pattern Recognition, pages 5762–5771. IEEE, 2016. [59] Kirill Fayn, Carolyn MacCann, Niko Tiliopoulos, and Paul J Silvia. Aesthetic emotions and aesthetic people: Openness predicts sensitivity to novelty in the experiences of interest and pleasure. *Frontiers in psychology*, 6:1877, 2015. [60] Mengjuan Fei, Wei Jiang, and Weijie Mao. Creating memorable video summaries that satisfy the user’s intention for taking the videos. *Neurocomputing*, 275:1911–1920, 2018. [61] Catrin Finkenauer, Rutger CME Engels, and Wim Meeus. Keeping secrets from parents: Advantages and disadvantages of secrecy in adolescence. *Journal of Youth and Adolescence*, 31(2):123–136, 2002. [62] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007. [63] Barbara L. Fredrickson. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions., volume 56. American Psychological Association, 2001. [64] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. [65] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. [66] Johan Galtung. Cultural violence. *Journal of peace research*, 27(3):291–305, 1990. [67] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41, 2016. [68] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In Hellenic conference on artificial intelligence, pages 91–100. Springer, 2010. [69] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. [70] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, 50(2):1–36, 2017. [71] Yu Gong, Weiqiang Wang, Shuqiang Jiang, Qingming Huang, and Wen Gao. Detecting violent scenes in movies by auditory and visual cues. In Pacific-Rim Conference on Multimedia, pages 317–326. Springer, 2008. [72] Shinichi Goto and Terumasa Aoki. TUDCL at mediaeval 2013 violent scenes detection: Training with multi-modal

features by MKL. In Proceedings of the MediaEval 2013 Workshop, 2013. [73] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. Visual interestingness in image sequences. In ACM international conference on Multimedia, pages 1017–1026. ACM, 2013. [74] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In Proceedings of the IEEE International Conference on Computer Vision, pages 1633–1640, 2013. 98 [75] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In European conference on computer vision, pages 505–520. Springer, 2014. [76] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3090–3098, 2015. [77] Michael Gygli and Mohammad Soleymani. Analyzing and predicting gif interestingness. In Proceedings of the 24th ACM international conference on Multimedia, pages 122–126, 2016. [78] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, et al. Can we measure beauty? computational evaluation of coral reef aesthetics. *PeerJ*, 3:e1390, 2015. [79] Raisa Halonen, Stina Westman, and Pirkko Oittinen. Naturalness and interestingness of test images for visual quality evaluation. In IS&T/SPIE Electronic Imaging, pages 78670Z–78670Z. International Society for Optics and Photonics, 2011. [80] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. Learning computational models of video memorability from fmri brain imaging. *IEEE transactions on cybernetics*, 45(8):1692–1703, 2014. [81] Alex Hanson, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry Davis. Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 0–0, 2018. [82] Judith M Harackiewicz, Jessi L Smith, and Stacy J Priniski. Interest matters: The importance of promoting interest in education. *Policy insights from the behavioral and brain sciences*, 3(2):220–227, 2016. [83] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 1–6. IEEE, 2012. [84] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. [85] Alejandro Hernández-García, Fernando Fernández-Martínez, and Fernando Díaz-de María. Comparing visual descriptors and automatic rating strategies for video aesthetics prediction. *Signal Processing: Image Communication*, 47:280–288, 2016. 99 [86] Eckhard H Hess and James M Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350, 1960. [87] Suzanne Hidi and Valerie Anderson. Situational interest and its impact on reading and expository writing. *The role of interest in learning and development*, 11:213–214, 1992. [88] Suzanne Hidi and William Baird. Interestingness—a neglected variable in discourse processing. *Cognitive science*, 10(2):179–194, 1986. [89] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. [90] Liang-Chi Hsieh, Winston H Hsu, and Hao-Chuan Wang. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4309–4313. IEEE, 2014. [91] L Rowell Huesmann. The impact of electronic media violence: Scientific theory and research. *Journal of Adolescent health*, 41(6):S6–S13, 2007. [92] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In Advances in neural information processing systems, pages 2429–2437, 2011. [93] Phillip



Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[94] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011.

[95] C. E. Izard and B. P. Ackerman. Emotion-cognition relationships and human development. Lewis, Michael and Haviland-Jones, Jeannette M, *Handbook of emotions*, pages 253–264, 2010.

[96] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and predicting interestingness of videos. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[97] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 73–79. ACM, 2014.

[98] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168, 2015.

[99] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Predicting subjectivity in image aesthetics assessment. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSp)*, pages 1–6. IEEE, 2019.

[100] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[101] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE, 2006.

[102] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[103] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *Advances in neural information processing systems*, pages 296–304, 2012.

[104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[105] Bert Krages. *Photography: the art of composition*. Skyhorse Publishing, Inc., 2012.

[106] Bartosz Krawczyk, Leandro L Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[108] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[109] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. *International affective picture system (iaps): Instruction manual and affective ratings*. The center for research in psychophysiology, University of Florida, 1999.

[110] Jieun Lee and Ilyoo B Hong. Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity. *International Journal of Information Management*, 36(3):360–373, 2016.

[111] Roberto Leyva, Faiyaz Doctor, Alba G. Seco de Herrera, and Sohail Sahab. Multimodal deep features fusion for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019.

[112] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing*, 3(2):236–252, 2009.

[113] Xirong Li, Yujia Huo, Qin Jin, and Jieping Xu. Detecting violence in video using subclasses. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 586–590, 2016.

[114] Cynthia CS Liem. *Tud-mmc at mediaeval 2016: Predicting media*

interesting- ness task. In Working Notes Proceedings of the MediaEval 2016 Workshop., 2016. [115] Feng Liu, Yuzhen Niu, and Michael Gleicher. Using web photos for measuring video frame interestingness. In Twenty-First International Joint Conference on Artificial Intelligence, pages 2058–2063, 2009. [116] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia, pages 83–92, 2010. [117] Gwenaëlle Marquant, Claire-Hélène Demarty, Christel Chamaret, Joël Sirot, and Louis Chevallier. Interestingness prediction & its application to immersive content. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pages 1–6. IEEE, 2018. [118] Albert Mehrabian. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. 1980. [119] Shasha Mo, Jianwei Niu, Yiming Su, and Sajal K Das. A novel feature set for video emotion recognition. *Neurocomputing*, 291:11–20, 2018. [120] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2408–2415. IEEE, 2012. [121] Enrique Bermejo Nieves, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In International conference on Computer analysis of images and patterns, pages 332–339. Springer, 2011. [122] Pardis Noorzad and Bob L Sturm. Regression with sparse approximations of data. In European Signal Processing Conference (EUSIPCO), pages 674–678. IEEE, 2012. 102 [123] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. [124] Balaji Padmanabhan and Alexander Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999. [125] Jayneel Parekh, Harshvardhan Tibrewal, and Sanjeel Parekh. Deep pairwise classification and ranking for predicting media interestingness. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pages 428–433, 2018. [126] Obioma Pelka, Christophe M Friedrich, A García Seco de Herrera, and Henning Müller. Overview of the imageclefmed 2019 concept detection task. CLEF working notes, CEUR, 2019. [127] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros. Technicolor and INRIA/IRISA at mediaeval 2011: learning temporal modality integration with bayesian networks. In Proceedings of the MediaEval 2011 Workshop, 2011. [128] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros. Technicolor-inria team at the mediaeval 2013 violent scenes detection task. In Proceedings of the MediaEval 2013 Workshop, 2013. [129] Cédric Penet, Claire-Hélène Demarty, Mohammad Soleymani, Guillaume Gravier, and Patrick Gros. Technicolor/inria/imperial college london at the mediaeval 2012 violent scene detection task. In Working Notes Proceedings of the MediaEval 2012 Workshop., 2012. [130] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In IEEE Conference on Computer Vision and Pattern Recognition, pages 860–868. IEEE, 2015. [131] Reza Aditya Permadi, Septian Gilang Permana Putra, Cynthia Helmiriawan, and Cynthia CS Liem. Dut-mmsr at mediaeval 2017: Predicting media interestingness task. In Working Notes Proceedings of the MediaEval 2017 Workshop., 2017. [132] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980. [133] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and social psychology review*, 8(4):364–382, 2004. 103 [134] Alison Reboud, Ismail Harrando, Jorma Laaksonen,

Danny Francis, Raphaël Troncy, and Héctor Laria Mantecón. Combining textual and visual modeling for predicting media memorability. In Working Notes Proceedings of the MediaEval 2019., 2019. [135] Miriam Redi and Bernard Merialdo. Where is the interestingness? retrieving appealing video scenes by learning flickr-based graded judgments. In International Conference on Multimedia Retrieval, pages 1363–1364. ACM, 2012. [136] Ronald A Rensink, J Kevin O’Regan, and James J Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, 8(5):368–373, 1997. [137] Lior Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12):4046–4072, 2009. [138] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 18–33. Springer Berlin Heidelberg, 2011. [139] Michael S Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. Assembled: Searching for multi-stream neural connectivity in video architectures. arXiv preprint arXiv:1905.13209, 2019. [140] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. [141] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. Insiders and outsiders: Comparing urban impressions between population groups. In *ACM on International Conference on Multimedia Retrieval*, pages 65–71. ACM, 2017. [142] Samuel Felipe dos Santos and Jurandy Santos. Gibis at mediaeval 2019: Predicting media memorability task. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019. [143] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. Who’s afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM international conference on Multimedia*, pages 311–320. ACM, 2015. [144] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *AAAI Conference on Web and Social Media*, 2015. 104 [145] Jan Schlüter, Bogdan Ionescu, Ionuț Mironică, and Markus Schedl. ARF @ mediaeval 2012: An uninformed approach to violence detection in hollywood movies. In *Proceedings of the MediaEval 2012 Workshop*, 2012. [146] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2009. [147] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. Umons at mediaeval 2015 affective impact of movies task including violent scenes detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*, 2015. [148] Amanda JC Sharkey. Types of multinet system. In *International Workshop on Multiple Classifier Systems*, pages 108–117. Springer, 2002. [149] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739, 2017. [150] Yuesong Shen, Claire-Hélène Demarty, and Ngoc Q. K. Duong. Technicolor@mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016. [151] Roger N Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 6(1):156–163, 1967. [152] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable? a deep style transfer

approach. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pages 322–329, 2017. [153] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89–102, 2005. [154] Paul J Silvia. Exploring the psychology of interest. Oxford University Press, 2006. [155] Paul J Silvia. Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3(1):48, 2009. [156] Paul J Silvia and John B Warburton. Positive and negative affect: Bridging states and traits. *Comprehensive handbook of personality and psychopathology*, 1:268–284, 2006. 105 [157] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. [158] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. The mediaeval 2015 affective impact of movies task. In Working Notes Proceedings of the MediaEval 2015 Workshop., 2015. [159] Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, Claire-Hélène Demarty, et al. The mediaeval 2014 affect task: Violent scenes detection. In Working Notes Proceedings of the MediaEval 2014 Workshop., 2014. [160] Mohammad Soleymani. The quest for visual interest. In Proceedings of the 23rd ACM international conference on Multimedia, pages 919–922, 2015. [161] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. [162] Liviu-Daniel Ştefan, Mihai Gabriel Constantin, and Bogdan Ionescu. System fusion with deep ensembles. In Proceedings of the 2020 International Conference on Multimedia Retrieval, pages 256–260, 2020. [163] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1102–1111, 2020. [164] Swathikiran Sudhakaran and Oswald Lanz. 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), pages 1–6. IEEE, 2017. [165] Jennifer J Sun, Ting Liu, and Gautam Prasad. Gla in mediaeval 2018 emotional impact of movies task. arXiv preprint arXiv:1911.12361, 2019. [166] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016. [167] Masaki Takahashi and Masanori Sano. Nhk where is beauty? grand challenge. In ACM Multimedia challenge, 2013. [168] Chun Chet Tan and Chong-Wah Ngo. The vireo team at mediaeval 2013: Violent scenes detection by mid-level concepts learnt from youtube. In Proceedings of the MediaEval 2013 Workshop, 2013. 106 [169] Silvan S Tomkins. Affect, imagery, consciousness: Vol. i. the positive affects. 1962. [170] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In European conference on computer vision, pages 776–789. Springer, 2010. [171] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015. [172] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. Predicting media memorability using deep features with attention and recurrent network. In Working Notes Proceedings of the MediaEval 2019 Workshop., 2019. [173] Samuel A Turner Jr and Paul J Silvia. Must interesting things be pleasant? a test of competing appraisal structures. *Emotion*, 6(4):670, 2006. [174] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of experimental psychology: General*, 123(4):394–409, 1994. [175] Joost Van De Weijer, Cordelia

Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009. [176] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Eth-cvl@ mediaeval 2016: Textual-visual embeddings and video2gif for video interestingness. In *Working Notes Proceedings of the MediaEval 2016 Workshop.*, 2016. [177] Alexander Viola and Sejong Yoon. A hybrid approach for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019. [178] Lijie Wang, Xueting Wang, and Toshihiko Yamasaki. Image aesthetics prediction using multiple patches preserving the original aspect ratio of contents. *arXiv preprint arXiv:2007.02268*, 2020. [179] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. [180] Shuai Wang, Shizhe Chen, Jinming Zhao, and Qin Jin. 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*, pages 55–61, 2018. 107 [181] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. Ruc at mediaeval 2019: Video memorability prediction based on visual textual and concept related features. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*, 2019. [182] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. [183] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270, 2018. [184] Ying Xu, Yi Wang, Huaixuan Zhang, and Yong Jiang. Spatial attentive image aesthetic assessment. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. [185] Wojtek Zajdel, Johannes D Krijnders, Tjeerd Andringa, and Dariu M Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *2007 IEEE conference on advanced video and signal based surveillance*, pages 200–205. IEEE, 2007. [186] Nick Zangwill. Aesthetic judgment. In E. N. Zalta, editor, *The Stanford encyclopedia of philosophy*. Fall 2007 ed. edition, 2003. [187] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3320. IEEE, 2011. [188] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM international conference on Multimedia*, pages 47–56. ACM, 2014. 108 3 4 9 16 18 19 22 24 25 27 28 30 34 35 38 40 41 42 43 47 55 58 59 60 66 70 73 80 81 88 89 91 92