UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicații și Tehnologia Informației

Decizie nr. 569 **din** 25-09-2020

REZUMAT TEZĂ DE DOCTORAT

Ing. Mihai Gabriel CONSTANTIN

AUTOMATIC ANALYSIS OF THE VISUAL IMPACT OF MULTIMEDIA DATA

ANALIZA AUTOMATĂ A IMPACTULUI VIZUAL AL DATELOR MULTIMEDIA

COMISIA DE DOCTORAT

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

BUCUREŞTI 2021

Chapter 1: Introduction

1.1 Domain of the thesis

This thesis covers several theoretical aspects and state-of-the-art methods on the automatic analysis of the visual impact of multimedia data. While more traditional computer vision tasks address problems that have objective ground-truth values that most annotators would agree with, recent research directions tend to study subjective concepts, like interestingness, aesthetics, violence, etc. In these cases ground truth may depend on a large set of human-related factors, including but not limited to personal preference, cultural background, and current psychological state.

1.2 Motivation of the thesis

This thesis aims to contribute to the understanding of a series of subjective concepts, discover and underline some good practices for the research of such concepts, and develop computer vision methods that can accurately predict them. While the 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. There is a growing demand for these methods, mostly driven by social media, media sharing, advertising, and media archiving platforms, which 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.

Interest and support from the industry for this type of research is so far demonstrated through several online applications, that represent parts and features of larger platforms, and the organization of benchmarking competitions that aid both the research community and the industry. Flickr social interestingness application[1] and Google Photos summary creation[2] represent some of the most popular industrial applications, while InterDigital[3] created several datasets and benchmarking competitions on interestingness, violence and memorability.

[1] https://www.flickr.com/

[2] https://www.google.com/photos/about/

[3] 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 [1]. Thus, the need arises for systems that can automatically process data, and filter or create suggestion lists according to human 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 concepts does make their prediction and classification one of the more challenging tasks in computer vision today.

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

The first concept analyzed in this thesis, *interestingness* has been defined as a primary factor for motivation and an important behavioral incentive for humans [2, 3]. Hidi and Anderson [4] propose that the appeal of an activity can be more important in generating interest than personal preferences. *Aesthetic value* is defined as a branch of philosophy that studies the appeal and beauty of compositions [5]. From a visual standpoint, *memorability* is defined as an intrinsic property of visual samples that measures how likely subjects are to remember the images and videos that are presented to them. Many authors use short- and long-term memorability [6] in defining the amount of time that the subject can retain the memorized information. Regarding *violence*, some authors [7] propose both a subjective definition (visual samples "which one would not let an eight years old child see, because they contain physical violence") and an objective definition ("physical violence or accident resulting in human injury or pain") for violence. Finally, *affective value and emotions,* is defined as the ability of media samples to induce certain emotional responses in subjects [8]. These emotions are mainly described in two ways: either in a mathematical 2D or 3D space with arousal, valence and dominance as the main features [9], or in a categorical space, containing emotions like anger, fear, joy, surprise, etc [10].

2.2 Human understanding of the subjective properties of multimedia data

Studying how humans perceive and interact with multimedia data is vital for this domain, as it creates a strong background that aids scientists in the computer vision domain by providing a set of principles that can be developed upon.

Interestingness. Berlyne [11] and Silvia [3] identify several factors that influence general interest, including novelty, complexity, uncertainty and conflict, however, as shown in [12] these relationships can be quite complex and non-linear. From an evolutionary perspective, Izard and Ackerman [13] conclude that interest allows humans to explore, learn and engage their environment. *Aesthetic value.* Reber et al. [14] propose that "goodness of form, symmetry and figure-ground contrast" are qualities that an item must have in order to be deemed aesthetically pleasing. Authors proposed a set of "rules of photography" that must be taken into account when analyzing the aesthetic quality of visual samples [15]. *Memorability.* As most of the studies in this domain show [16], the human mind has an impressive and perhaps unexpected capacity for remembering visual data. *Violence.* Arendt [17] 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 [18] attempts to study it from a cultural perspective, noticing the intra-cultural difference of perception of violence. *Affective value and emotions.* Emotions have been studied from many perspectives, ranging from color theory [19] to an educational perspective [20].

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 properties. 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 inherent subjectivity of the analyzed multimedia data.

Perhaps some of the most impactful datasets are represented by works that analyze more than one concept. One example from this category is the visInterest [21] dataset that has interestingness as the main concept, but also includes concepts that are theorized to influence interest, such as coping potential, complexity and arousal.

Other works are built around the idea of a *common evaluation benchmark*, and provide not only data and annotations, but also a set of descriptors, metrics, data splits, creating an environment where method performance can be correctly performed. Such datasets are built around interestingness [22], memorability [6] and violence [23].

2.4 Computational approaches

2.4.1 Interestingness

While many computational approaches have been tested for the prediction of media interestingness, so far deep neural networks have not achieved optimal performance. While some authors attempt to use related concepts for predicting interestingness, like novelty and aesthetics [24], usually through the use of traditional visual features, others use the features directly for predicting interestingness [25] .The MediaEval Predicting Media Interestingness [22, 26] task gave the opportunity to test several systems in the same setup with regards to dataset, training / testing splits and metrics.

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 [15, 27, 28]. 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.

2.4.3 Memorability

Early methods for memorability prediction [29] merge human studies with computer vision methods for image classification, using conclusions drawn from the former in designing the latter. More modern approaches fully use the power of deep neural networks. For example, visual attention mechanisms and LSTM layers [30] are deployed in a ResNet-based convolutional architecture by Fajtl et al. [31].

2.4.4 Violence

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. Some examples include the use of traditional motion features [32], flow-vector magnitudes [33] or LSTM-based approaches [34].

2.4.5 Affective value and emotions

A large body of literature is dedicated to emotional content prediction. Zhao et al. [35] 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. Specialized sentiment features [36] and arousal features [37] are also used for indicating emotional content.

Chapter 3: Personal contributions

3.1 Datasets and evaluation

This chapter presents my contributions to the creation of several publicly available datasets including: (i) Interestingness10k [38], designed for the prediction of image and video interestingness; (ii) VSD96 [39], a video dataset for violent scenes detection; (iii) the MediaEval 2019 Predicting Media Memorability [6] a dataset composed of short videos that are annotated with short-term and long-term memorability values; and finally (iv) the MMTF-14k [40], a dataset for movie recommendation.

3.1.1 Interestingness prediction

The Interestingness10k [38] dataset is a publicly available[4] dataset and a common evaluation framework, designed for the prediction of image and video interestingness, validated and tested during the MediaEval 2016 and 2017 Predicting Media Interestingness tasks. My main contributions to this 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. The dataset is composed of image and video samples extracted from Creative Commons[5] licensed Hollywood-like movies, split into 7,396 samples in the development set and 2,192 samples in the testing set, in the latest version of the dataset.

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

[5] https://creativecommons.org/

For the *overall performance analysis,* we gathered all the participant systems from the MediaEval competitions and analyzed trends and improvements. The most important observation in this case is that system performance improved between the two editions of the task, by 25.75% for the image task and 22.75% for the video task. It is also interesting to note that, while human annotator performance is above the automatic prediction system performance, humans never reach a near-perfect performance either, with their results being under MAP = 0.7.

The *feature-level analysis* shows that six main feature types are employed by participants: visual, audio, motion, deep learning-based, conceptual and textual. Many systems employ more than one feature type in various fusion schemes, creating 18 combinations of these features. On average, for the image task, deep features have the better performance (MAP = 0.2297), while on the video task traditional visual features perform better (MAP = 0.1798).

The analysis of *generalization capabilities* shows some interesting conclusions regarding participant systems. For example, for the image task, deep learning systems that undergo a pre-training step even on an uncorrelated dataset show better performance than systems that do not use pre-training. Also, there is a correlation of results (calculated via Pearson's Correlation = 0.546) between the performances of similar image and video prediction systems, indicating that adapting image predictors to videos may represent a good starting point. Finally, system performance on longer videos was superior to performance on short videos (MAP@10 = 0.751 vs. 0.0562), indicating that longer videos create a larger separation between the two classes.

Finally, we create a set of recommendations with regards to system performance that includes the following ideas:

- 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 [41], can have good performances, however they do not surpass the current state-of-the-art;
- upsampling has a positive effect on system performance, as shown in [42];
- system performance may benefit from pre-training on external data [43].

3.1.2 Violence prediction

The VSD96 dataset [39] is a publicly available dataset[6] [7] and a common evaluation framework designed for the detection of violent scenes in Hollywood-like and

[6] Data for 2011-2014 available at: https://www.interdigital.com/data_sets/violent-scenes-dataset

[7] Data for 2015 available at: http://liris-accede.ec-lyon.fr/

YouTube[8] movies. Versions of this dataset were used during the 2011-2015 MediaEval Violent Scenes Detection task. 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.

The *overall system analysis* shows that overall, the performance of participating systems has improved, reaching a MAP of 0.51 for shot-level detection using the objective definition of violence. Also encouraging are the good results recorded on the YouTube generalization part of the dataset, given that systems were not trained on that particular type of video data, thus showing a good understanding of the general concept of violence.

The *analysis of employed features* shows that participants mainly employed four types of features: visual, audio, concept and deep learning features. While the first three are used throughout the editions of the competitions, deep learning features start showing their popularity in the 2014 and especially in the 2015 edition of the competition. Overall 12 combinations of these modalities are used by participants. Furthermore, with regards to multimodal systems, four categories stand out, obtaining top results in certain subtasks: (i) visual and audio, (ii) audio and conceptual, (iii) visual, audio and conceptual, and (iv) visual, audio and deep. Finally, late fusion systems achieve a better MAP performance than early or unimodal systems.

3.1.3 Memorability prediction

The MediaEval 2019 Predicting Media Memorability dataset [6], 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. The dataset consists of 10,000 short soundless videos, split between the development set (80% of the data) and testing set (20% of the data). Overall this edition of the competition shows significant improvements over earlier editions in system performance.

3.1.4 Content Recommendation

The MMTF-14K [40] is a publicly available dataset[9] 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.

[8] www.youtube.com

[9] https://zenodo.org/record/1225406.Xw830s8zaXw

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 [44] as well as learning systems based on the use of aesthetic features and late fusion [45, 46]. 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.

3.2.2 SVM-based learning systems

This approach consists of three phases, as shown in Figure 3.2.1. The first stage involves the extraction of a set of traditional video features, followed by a stage that fuses the features in various combinations and ending with a SVM-based learning method.

A set of seven features is extracted for each of the images and videos in the dataset, including: (i) color histogram in the HSV space, (ii) dense SIFT transform, (iii) LBP, (iv) HoG, (v) GIST, (vi) features extracted from the fc7 and prob layers of the AlexNet architecture [47], (vii) the color naming histogram [48].

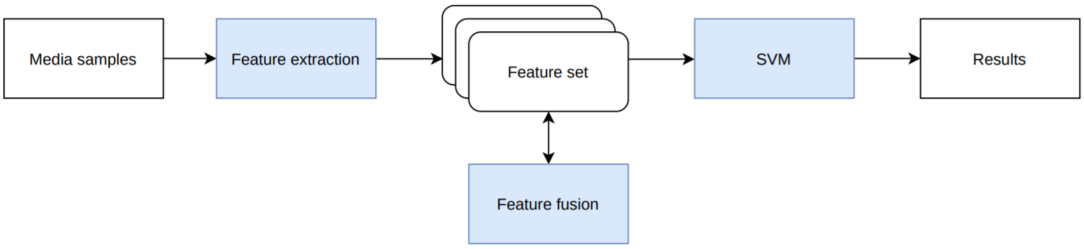

Figure 3.2.1 The diagram of the proposed SVM-based method. The three main stages (Feature extraction, Feature fusion and SVM) are highlighted in blue.

For image prediction, each feature is extracted individually and treated as a vector of floating-point values, while for video prediction, individual frames are extracted and then aggregated at video-level by averaging the frame vectors. Fusion is realized by concatenating various vectors of features.

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 $k \in [1,..., 10]$;
- gamma coefficient (denoted γ) with values of 2^k, where $k \in [1,..., 6]$;

while for the RBF kernels the following parameters are tested:

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

Experimental setup. The various combinations of features and SVM learners are tested in the context of the MediaEval 2016 Predicting Media Interestingness Task [26].

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

Table 3.2.1 System results on the testset, showing the best results of the submitted systems, compared with average and best performing systems at the MediaEval competition.

Table 3.2.1 presents the results on the testset of our top performing systems. 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).

3.2.3 Aesthetic features and late fusion learning systems

Given the previous results [44] 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 [38], aesthetic appeal and interestingness are quite often studied together. We decide to extract a set of aesthetic-based features, developed in [15, 27, 28] and use these features for the prediction of media interestingness. We test this approach on the MediaEval 2016 [26] and 2017 [22] Predicting Media Interestingness Task datasets, publishing this approach in two papers [45, 46].

Figure 3.2.2 presents a diagram of this approach. At this level, the difference between this and our previous methods is represented by the appearance of a late fusion fourth stage.

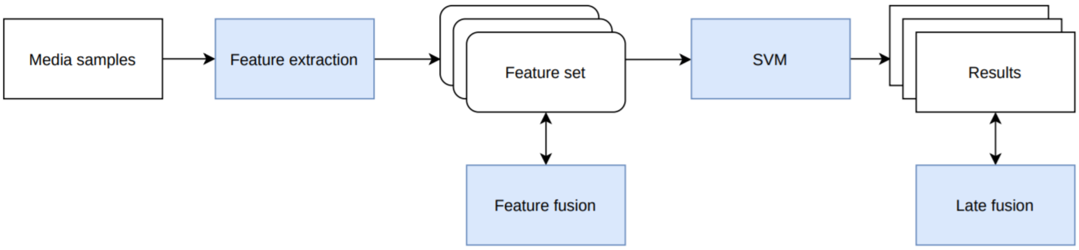

Figure 3.2.2 The diagram of the proposed SVM-based method. The three main stages (Feature extraction, Feature fusion, SVM and Late fusion) are highlighted in blue.

Our approach uses a set SVM classifiers with polynomial, RBF, and linear kernels, and augmented via early and late fusion.

With regards to the aesthetic descriptors, three main groups of features are used in this work, as described in [49]: (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. Furthermore, the same SVM training parameters are used as in the previous experiments.

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.

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

Table 3.2.1 System results on the testset, showing the best results of the submitted systems, compared with average and best performing systems at the MediaEval 2016 competition.

Table 3.2.1 shows the best results on the MediaEval 2016 competition, with both early and late fusion systems performing above the best state-of-the-art system. It is also interesting to note that the top inducer has also performed above the average results from the MediaEval competition. Top performance was achieved by a CombMax late fusion system.

Task	System	MAP testset	MAP@10 testset
image	ME top	0.3075	0.1385
	ME avg	0.2402	0.0876
	CombMean (aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.1873	0.5555

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

Table 3.2.2 System results on the testset, showing the best results of the submitted systems, compared with average and best performing systems at the MediaEval 2017 competition.

Table 3.2.2 shows the best results for the MediaEval 2017 competition, where again the best performing system is a late fusion approach. While this time the proposed approaches did not surpass the top performers in the MediaEval competition, for the video task performance was above average.

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 surveillance videos. This approach employs a ConvLSTM [50] structure that processes visual features created by processing video frame differences with a VGG [51] network. Experiments with this approach are validated on two datasets: the MediaEval 2015 Violent Scene Detection dataset [23] and the VIF dataset [33].

3.3.3 Temporal deep learning systems

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. 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 in [52], under the assumption that the feature extracting networks will 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 [51], which will encode a set of features for each pair of frame differences. In the final phase, ConvLSTM [50] 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 × 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 3.3.1.

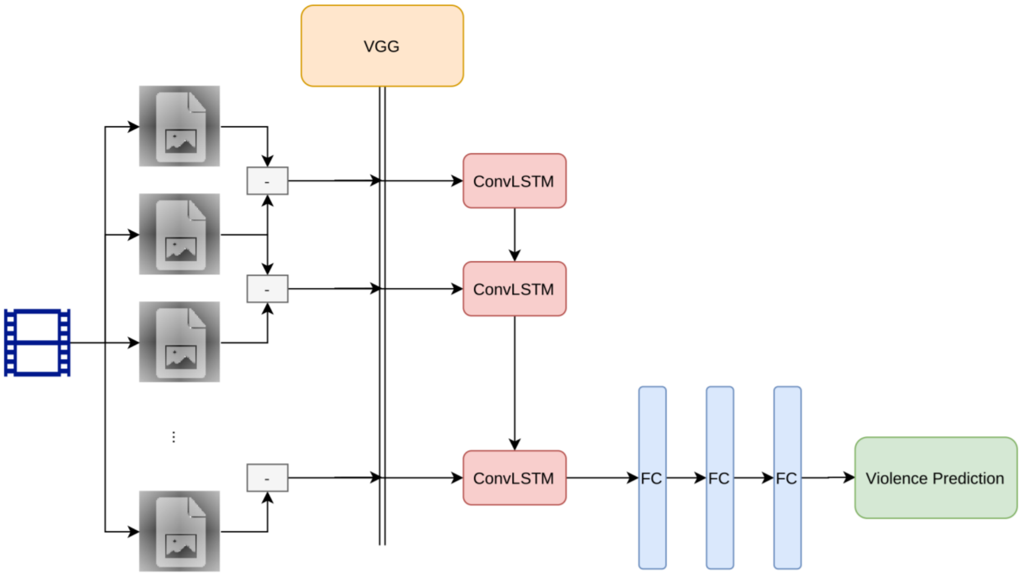

Figure 3.3.1 The diagram of the proposed temporal deep neural network approach.

Experimental results are presented in Table 3.3.1, 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, 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.

Method	Window config.	VSD2015 (MAP)	VIF (Acc)
SOA	-	0.296	0.863
Proposed system	30	0.271	0.89

Table 3.3.1 System results on the two datasets, compared with the current top state-of-the-art performance on each 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 [53] 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.

3.4.2 Action-based deep learning systems

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 [30], included in their architectures in order to produce better results on temporal data. 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.

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 [54] is fine-tuned on the memorability data. At the same time, for the action recognition DNNs the TSN [55] and I3D [56] networks are used as feature extractors and augmented with the C3D features [57] 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 3.4.1.

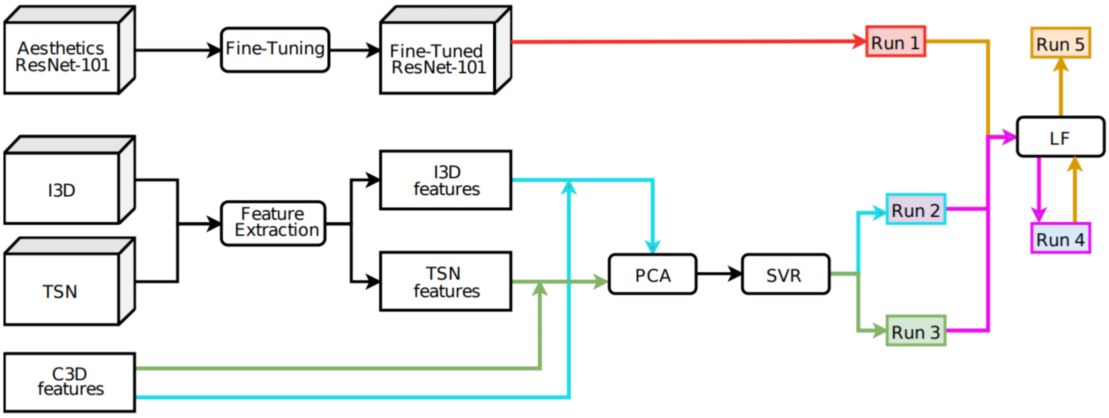

Figure 3.4.1 The diagram of the proposed solution. We represent the aesthetic-based network (ResNet-101) and action recognition networks (I3D, TSN, and the organizer provided C3D), their fine-tuning or extraction process and learning process, and the final late fusion (LF) stage. The components of the five individual runs submitted to the MediaEval Predicting Media Memorability task are also represented (Run 1 - Run 5)

The aesthetic based architecture is described in Kang et al. [58] . We extract the “Mixed_5” layer and use it as a feature from the I3D model, trained on the Kinetics dataset [59], while the “Inception_5”' layer is extracted from the TSN model, trained on the UCF101 dataset [60]. We perform preliminary tests 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, the three late fusion schemes we employ are CombMax, CombMin, and CombMean.

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

Table 3.4.1 System results on the Predicting Media Memorability dataset, compared against the top results.

The final results on the testset, shown in Table 3.4.1, 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 ρ= 0.477 for short-term memorability and ρ = 0.232 for long-term memorability.

3.5 Late fusion with deep ensembling systems

3.5.1 Introduction

In this chapter, we present the contributions to the creation of deep ensembling systems. Our works [61, 62] and [38][10] 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.

[10] Paper under revision

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 and memorability 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 [63], video memorability [64], and emotional content analysis [65], but also domains that do not deal with such subjective concepts, examples here including the classification of human actions in videos [41].

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 [66] and Gradient Boosting [67].

3.5.4 Proposed method

The proposed DeepFusion approach deploys several types of DNN that take as input the set of inducer outputs 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 3.5.1. 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, (ii)the number of neurons per layer, 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.

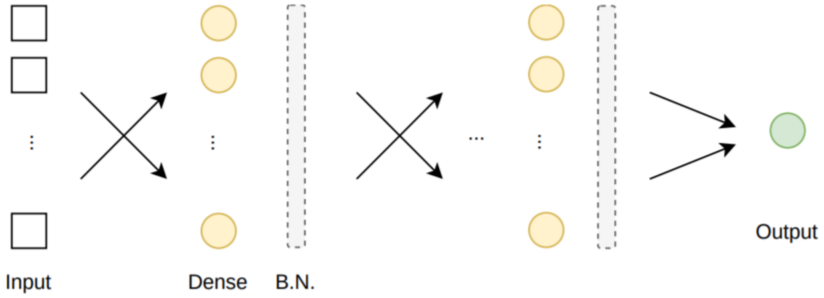

Figure 3.5.1 DeepFusion dense network architecture (DF-Dense): variable number of layers, number of neurons per layer and the presence or absence of Batch Normalization (BN) layers.

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, 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.

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

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

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

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.

Dense networks with convolutional layers

A general presentation of the employed convolutional architecture is depicted in Figure 3.5.2. After processing the input and transforming it into a $tr2D$form, this input is fed into a convolutional layer. Given the 3 × 3 padding of each element of the original input, we also choose to use a 3 × 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 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.

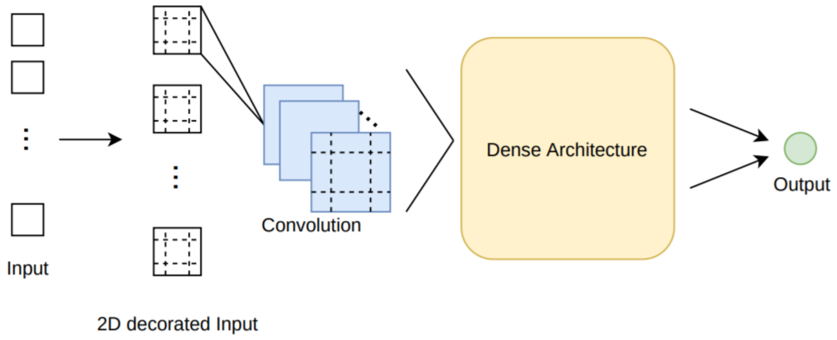

Figure 3.5.2 DeepFusion convolutional network architecture (DF-Conv). Represented here are the input processing stage, convolutional filters and trailing Dense Architecture.

Dense networks with cross-space-fusion layers

Finally, we introduce the Cross-Space-Fusion (CSF) layer, whose general design is presented in Figure 3.5.3. This layer takes the 3D $tr3D$ array and, for each group of centroids ($tr3Dc$, $tr3Dr$) learns a set of weights α, β, that process the 3D input as follows:

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

The number of parameters used by the CSF layer per each centroid pair is 16, thus generating 16 × *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 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.

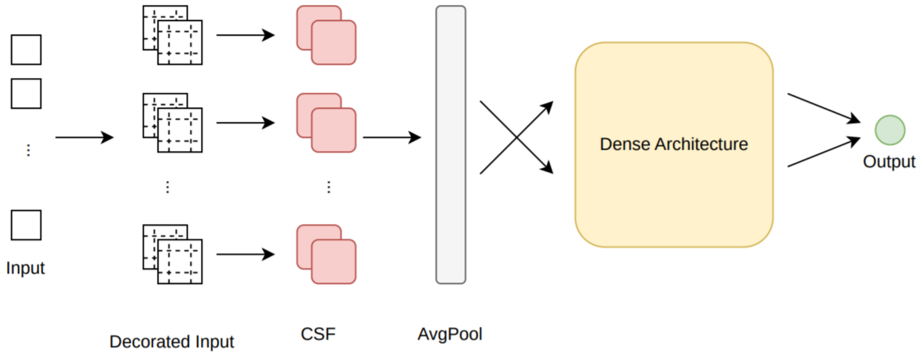

Figure 3.5.3 DeepFusion Cross-Space-Fusion network architecture (DF-CSF). Represented here are the input decoration stage, CSF processing layer, Average Pooling layer and trailing Dense Architecture.

Training protocol

We propose several essential steps in developing this late fusion approach. The first step consists of gathering all the individual 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”.

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. We deployed our methods on the following datasets: MediaEval 2017 Predicting Media Interestingness [22] split into an image subtask (denoted INT2017.Image) and a video subtask (INT2017.Video), MediaEval 2015 Violent Scenes Detection [23] (VSD2015.Video), MediaEval 2018 Predicting Emotional Impact of Movies [68] split into an arousal (Aro2018.Video), valence (Val2018.Video) and fear detection (Fear2018.Video), and finally the ImageCLEFmed 2019 Concept Detection [69] (Capt2019.Image).

3.5.6 Experimental results

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

Table 2.5.1 Results on the INT2017.Image, INT2017.Video and VSD2015.Video datasets for the dense and convolutional architectures.

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

Table 2.5.2 Results on the Aro2018.Video, Val2018.Video, Fear2018.Video and Capt2019.Image datasets for the dense and CSF architectures.

As shown in Tables 2.5.1 and 2.5.2, the results for these proposed architectures clearly surpassed not only the current state-of-the-art (ME top, SOA top), but also a set of traditional embedding methods (Emb). These results represent a significant improvement over the state-of-the-art systems, even going up to 200.9% improvement in the case of INT2017.Video.

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:

- *Book chapters*

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

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

- *Journals*

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

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

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

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

- *Conferences*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In (C2) 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 (C3) and (C5) 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 datsets, 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 (J1), (C7), (C9), (C10) 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.

(J2) currently represents, to the best of my knowledge, 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 (C11) I was the main organizer of the MediaEval 2019 Predicting Media Memorability task, with contributions in helping MediaEval participants, evaluating submitted systems and theorizing general trends with regards to best practices.

In (C12) 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.

(J3) 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 datset 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.

(J4), 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.

(C15) 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.

(C16) presents another set of novel approaches with regards to ensembling systems. While keeping the DNN-based architecture approach, the novelty of this paper is represented by the introduction of a DNN layer specially designed for this type of task, the Cross-Space-Fusion layer. My main contributions to this paper are represented by the creation of another input decoration method and by the creation and development of the CSF layer. Validation is carried out on a variety of tasks that cover various validation conditions: two-class regression (represented by the Arousal and Valence detection subtasks of the MediaEval 2018 Emotional Impact of Movies task), binary classification (represented by the Fear detection subtask of the MediaEval 2018 Emotional Impact of Movies task) and multi-label classification (represented by the ImageCLEF 2019 Medical Concept Detection task).

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 interestingess, 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.

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] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huber-man. Influence and passivity in social media. In Machine Learning and Knowledge Discovery in Databases, volume 6913, 18–33. Springer Berlin Heidelberg, 2011.

[2] Berlyne, D. E. (1949). Interest as a psychological concept. British Journal of Psychology, 39(4), 184.

[3] Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. Emotion, 5(1), 89.

[4] Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. The role of interest in learning and development, 11, 213-214.

[5] Zangwill, N. (2003). Aesthetic judgment.

[6] Constantin, M. G., Ionescu, B., Demarty, C. H., Duong, N. Q., Alameda-Pineda, X., & Sjöberg, M. (2019, October). Predicting Media Memorability Task at MediaEval 2019. In Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France.

[7] Demarty, C. H., Penet, C., Schedl, M., Bogdan, I., Quang, V. L., & Jiang, Y. G. (2013, October). The mediaeval 2013 affect task: violent scenes detection. In MediaEval 2013 Working Notes (p. 2).

[8] Cabanac, M. (2002). What is emotion?. Behavioural processes, 60(2), 69-83.

[9] Mehrabian, A. (1980). Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies (Vol. 2). Cambridge, MA: Oelgeschlager, Gunn & Hain.

[10] Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

[11] Berlyne, D. E. (1960). Conflict, arousal, and curiosity.

[12] Silvia, P. J. (2009). Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. Psychology of Aesthetics, Creativity, and the Arts, 3(1), 48.

[13] Izard, C. E. (1984). Emotion-cognition relationships and human. Emotions, cognition, and behavior, 17.

[14] Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?. Personality and social psychology review, 8(4), 364-382.

[15] Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006, May). Studying aesthetics in photographic images using a computational approach. In European conference on computer vision (pp. 288-301). Springer, Berlin, Heidelberg.

[16] Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. Journal of verbal Learning and verbal Behavior, 6(1), 156-163.

[17] Arendt, H. (1970). On violence. Houghton Mifflin Harcourt.

[18] Galtung, J. (1990). Cultural violence. Journal of peace research, 27(3), 291-305.

[19] Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. Journal of experimental psychology: General, 123(4), 394.

[20] Chen, C. M., & Sun, Y. C. (2012). Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. Computers & Education, 59(4), 1273-1285.

[21] Soleymani, M. (2015, October). The quest for visual interest. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 919-922).

[22] Demarty, C. H., Sjöberg, M., Ionescu, B., Do, T. T., Gygli, M., & Duong, N. (2017, September). Mediaeval 2017 predicting media interestingness task. In MediaEval workshop.

[23] Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., ... & Chen, L. (2015, September). The MediaEval 2015 Affective Impact of Movies Task. In MediaEval.

[24] Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1633-1640).

[25] Jiang, Y. G., Wang, Y., Feng, R., Xue, X., Zheng, Y., & Yang, H. (2013, June). Understanding and predicting interestingness of videos. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 27, No. 1).

[26] Demarty, C. H., Sjoberg, M., Ionescu, B., Do, T. T., Wang, H., Duong, N. Q., & Lefebvre, F. (2016). Mediaeval 2016 predicting media interestingness task. In MediaEval 2016 Workshop.

[27] Ke, Y., Tang, X., & Jing, F. (2006, June). The design of high-level features for photo quality assessment. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 1, pp. 419-426). IEEE.

[28] Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. IEEE Journal of selected topics in Signal Processing, 3(2), 236-252.

[29] Parikh, D., Isola, P., Torralba, A., & Oliva, A. (2012). Understanding the intrinsic memorability of images. Journal of Vision, 12(9), 1082-1082.

[30] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[31] Fajtl, J., Argyriou, V., Monekosso, D., & Remagnino, P. (2018). Amnet: Memorability estimation with attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6363-6372).

[32] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., & Theodoridis, S. (2010, May). Audio-visual fusion for detecting violent scenes in videos. In Hellenic conference on artificial intelligence (pp. 91-100). Springer, Berlin, Heidelberg.

[33] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012, June). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-6). IEEE.

[34] Hanson, A., Pnvr, K., Krishnagopal, S., & Davis, L. (2018). Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[35] Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T. S., & Sun, X. (2014, November). Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 47-56).

[36] Jou, B., Chen, T., Pappas, N., Redi, M., Topkara, M., & Chang, S. F. (2015, October). Visual affect around the world: A large-scale multilingual visual sentiment ontology. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 159-168).

[37] Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. Journal of experimental psychology: General, 123(4), 394.

[38] Constantin, M. G., Ştefan, L. D., Ionescu, B., Duong, N. Q., Demarty, C. H., & Sjöberg, M. (2021). Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision, 1-25.

[39] Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C. H., Sjoberg, M., Schedl, M., & Gravier, G. (2020). Affect in multimedia: Benchmarking violent scenes detection. IEEE Transactions on Affective Computing.

[40] Deldjoo, Y., Constantin, M. G., Ionescu, B., Schedl, M., & Cremonesi, P. (2018, June). MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 450-455).

[41] Sudhakaran, S., Escalera, S., & Lanz, O. (2020). Gate-shift networks for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1102-1111).

[42] Shen, Y, Demarty, C. H., Duong, N. Q. K. Technicolor@MediaEval 2016 Predicting Media Interestingness Task

[43] Vasudevan, A. B., Gygli, M., Volokitin, A., & Van Gool, L. (2016, October). ETH-CVL@ MediaEval 2016: Textual-Visual Embeddings and Video2GIF for Video Interestingness. In MediaEval.

[44] Constantin, M. G., Boteanu, B., & Ionescu, B. (2016, October). LAPI at MediaEval 2016 Predicting Media Interestingness Task. In MediaEval.

[45] Constantin, M. G., Boteanu, B. A., & Ionescu, B. (2017). LAPI at MediaEval 2017-Predicting Media Interestingness. In MediaEval.

[46] Constantin, M. G., & Ionescu, B. (2017, July). Content description for Predicting image Interestingness. In 2017 International Symposium on Signals, Circuits and Systems (ISSCS) (pp. 1-4). IEEE.

[47] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[48] Van De Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009). Learning color names for real-world applications. IEEE Transactions on Image Processing, 18(7), 1512-1523.

[49] Haas, A. F., Guibert, M., Foerschner, A., Calhoun, S., George, E., Hatay, M., ... & Rohwer, F. (2015). Can we measure beauty? Computational evaluation of coral reef aesthetics. PeerJ, 3, e1390.

[50] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. arXiv preprint arXiv:1506.04214.

[51] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[52] Sudhakaran, S., & Lanz, O. (2017, August). Learning to detect violent videos using convolutional long short-term memory. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE.

[53] Constantin, M. G., Kang, C., Dinu, G., Dufaux, F., Valenzise, G., & Ionescu, B. (2019, October). Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In MediaEval 2019 Workshop.

[54] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[55] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36). Springer, Cham.

[56] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[57] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[58] Kang, C., Valenzise, G., & Dufaux, F. (2019, September). Predicting Subjectivity in Image Aesthetics Assessment. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.

[59] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

[60] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

[61] Ştefan, L. D., Constantin, M. G., & Ionescu, B. (2020, June). System Fusion with Deep Ensembles. In Proceedings of the 2020 International Conference on Multimedia Retrieval (pp. 256-260).

[62] Constantin, M. G., Ştefan, L. D., & Ionescu, B. (2021, June). DeepFusion: Deep Ensembles for Domain Independent System Fusion. In International Conference on Multimedia Modeling (pp. 240-252). Springer, Cham.

[63] Wang, S., Chen, S., Zhao, J., & Jin, Q. (2018, October). Video interestingness prediction based on ranking model. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data (pp. 55-61).

[64] Azcona, D., Moreu, E., Hu, F., Ward, T. E., & Smeaton, A. F. (2020, September). Predicting media memorability using ensemble models. CEUR Workshop Proceedings.

[65] Sun, J. J., Liu, T., & Prasad, G. (2019). Gla in mediaeval 2018 emotional impact of movies task. arXiv preprint arXiv:1911.12361.

[66] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.

[67] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[68] Dellandréa, E., Huigsloot, M., Chen, L., Baveye, Y., Xiao, Z., & Sjöberg, M. (2018). The MediaEval 2018 Emotional Impact of Movies Task. In Multimedia Benchmark Workshop. CEUR.

[69] Pelka, O., Friedrich, C. M., Seco De Herrera, A. G., & Müller, H. (2019, July). Overview of the ImageCLEFmed 2019 concept detection task. CEUR Workshop Proceedings.