Noname manuscript No.
(will be inserted by the editor)

Visual Interestingness Prediction: A Benchmark Framework and Literature Review

Received: date / Accepted: date

Abstract In this paper, we report on the creation of a publicly available, common evaluation framework for image and video visual interestingness prediction. We propose a robust data set, the Interestingness10k, with 9,831 images and more than 4 hours of video, interestigness scores determined based on more than 1M pair-wise annotations of 800 trusted annotators, some pre-computed multi-modal descriptors, and 192 system output results as baselines. The data were validated extensively during the 2016-2017 MediaEval benchmark campaigns. We provide an in-depth analysis of the crucial components of visual interestingness prediction algorithms by reviewing the capabilities and the evolution of the MediaEval benchmark systems, as well as of prominent systems from the literature. We discuss overall trends, influence of the employed features and techniques, generalization capabilities and the reliability of results. We also discuss the possibility of going beyond state-of-the-art performance via an automatic, ad-hoc system fusion, and propose a deep MLP-based architecture that outperforms the current state-of-the-art systems by a large margin. Finally, we provide the most important lessons learned and insights gained.

M.G. Constantin, L.-D. Ştefan, B. Ionescu,
University Politehnica of Bucharest, Romania
E-mail: mgconstantin@imag.pub.ro, lstefan@imag.pub.ro, bogdan.ionescu@upb.ro

Q.-K.-N. Duong, C.-H. Demarty,
Technicolor, France
E-mail: Quang-Khanh-Ngoc.Duong@technicolor.com, claire-helene.demarty@technicolor.com

M. Sjöberg,
CSC - IT Center for Science Ltd., Finland
E-mail: mats.sjoberg@csc.fi

Keywords visual interestingness prediction · Interestingness10k data set · benchmarking · literature survey · late fusion techniques

1 Introduction

Recent advances in automatic analysis of multimedia information go beyond the annotation and prediction of concrete, tangible and objective concepts, such as the presence of specific objects or scene understanding. Motivated by the richness of computer applications where human interaction is central, researchers now also concentrate on the prediction of *subjective concepts*, related to *human behaviour* and *perception*, such as visual memorability, Squalli-Houssaini et al. [1], induced emotions, Mo et al. [2], or visual aesthetics, Carballal et al. [3].

When addressing human reactions and perception assessment of multimedia content, an important role is played by the person himself, personal preferences, individual personality, cultural backgrounds and many more subjective factors. This is an additional challenge to devising automatic machine learning algorithms, as it requires ground truth data specifically adapted to this human-oriented task.

In this work, we address this challenge and discuss resources and approaches for one of the most popular subjective concepts of visual information, namely *visual interestingness*, Constantin et al. [4]. *Interestingness* has been defined and studied for some time, starting with Berlyne's works in psychology [5], who classifies *interest* as a defining factor for human motivation and behaviour. Later, Berlyne [6, 7] identifies factors that induce or influence interest, including *novelty*, *complexity*, *uncertainty* and *conflict*. The high degree of *subjectiv-*

ity associated with interestingness is visible from some of its definitions, a crucial role in determining interestingness being assigned to the observer. For example, *situational interest* is defined by Hidi and Anderson [8] as "the appealing effect of an activity or learning task on an individual". Chamaret et al. [9] define interestingness as "the quantification of the ability of an image to induce interest in a user".

In some psychological studies, interestingness has been considered as an *emotion*, Silvia [10,11], and included in the *knowledge emotions* category that is related to the *comprehension* process. Interest has been shown to be a product of two appraisal structures: *novelty-complexity* (interest shown for new and complex events) and *coping potential* (the ability to understand an event). Further studies have also revealed subjective differences between the perception of interestingness, based on personality traits, e.g., subjects that had high values for their *openness* were more influenced by the *novelty-complexity* appraisal structure, McCrae [12].

In automated, computational approaches, the concept of *interestingness* is projected in two perspectives, Constantin et al. [4]: *visual interestingness*, which is related to the aforementioned definitions, and *social interestingness*, which is related to social media concepts such as *popularity*, *virality*, number of likes on social platforms, shares, etc. These concepts, although they may seem correlated, depending on the use case and data, proved to be, in fact, weakly correlated at best, typically negatively correlated, Hsieh et al. [13]. Items with high impact on social networks are not necessarily interesting from a visual perspective.

In this context, this work focuses on the concept of *visual interestingness* and proposes a *publicly available*, *common evaluation framework*, for the prediction of image and video visual interestingness. Proposed resources include large annotated data (the Interestingness10k data set) and evaluation protocols, as well as an in-depth study of benchmark and state-of-the-art approaches, with the objective of providing relevant baselines for a complete practitioner's guide. To disambiguate the information need, we adopt a real-world, Video on Demand (VOD), use case scenario, employed by Technicolor[1]. A computational system should be capable of automatically selecting movie images/parts which are considered to be the most interesting ones for the underlying movie, Demarty et al. [14]. The proposed resources have been validated during the 2016 and 2017 MediaEval Benchmarking Initiative for Multimedia Evaluation[2].

We strongly believe that this type of overview contribution that creates useful insights into its field has a significant impact and helps shape the research directions. We follow the best practices from the literature, like the evolution of PASCAL Visual Object Classes data set[3] in Everingham et al. [15], ILSVRC benchmark[4] in Russakovsky et al. [16], TRECVid[5] shot boundary detection track in Smeaton et al. [17], TRECVid content-based video copy detection benchmark in Awad et al. [18], ImageCLEF[6] automatic medical annotation data sets in Deselaers et al. [19], multimodal person discovery in broadcast TV benchmark in Poignant et al. [20], ImageCLEF biomedical image retrieval systems in Kalpathy-Cramer et al. [21].

Some of the most important insights to takeaway from our study can be summarized with the following: (i) Interestingness entails a high degree of annotator subjectivity; (ii) What is interesting in an image? analysis of annotator data reveals some specific patterns such as colored and aesthetic frames, and presence of people; (iii) System performance for prediction is much lower than for more objective tasks, such as object detection or scene classification. Even humans, while significantly surpassing machine performance, do not achieve perfect prediction; (iv) Current state-of-the-art deep neural networks, while achieving good performance, they are not the top prediction performers; (v) What deep neural networks learn? Grad-CAM analysis shows an explicit focus on the main subject, but also on the area around. The presence of people triggers activation also around the faces; (vi) Late fusion and ensemble systems represent a good option with implicit higher performance than single systems of any type.

The remainder of the article is structured as follows. Section 2 presents the state of the art and positions our contribution. Section 3 describes the proposed data set, including the annotation protocol. Section 4 presents the recommended evaluation protocol. Section 5 presents an in-depth analysis of benchmark and state-of-the-art systems: overall capabilities, employed descriptors, prediction methods, generalization capabilities, and reliability analysis. Section 6 investigates the performance of several state-of-the-art deep neural networks on the proposed data. Section 7 discusses the possibility of boosting performance by building an ad-hoc system on top of existing baselines and proposes a deep MLP-based solution. Section 8 concludes the paper and discusses future perspectives.

[1] https://www.technicolor.com/
[2] http://www.multimediaeval.org/
[3] http://host.robots.ox.ac.uk/pascal/VOC/
[4] http://www.image-net.org/challenges/LSVRC/
[5] https://trecvid.nist.gov/
[6] https://www.imageclef.org/

2 Previous work

We review the relevant literature on the resources available to benchmark and develop visual interestingness prediction algorithms. For a comprehensive study of computational approaches for interestingness prediction, the reader is referred to our previous contribution, Constantin et al. [4]. Interestingness data sets have been created with the goal of predicting either image or video interestingness. A summary is presented in Table 2.

For instance, the Scene categories data set created by Gygli et al. [22] is built on top of the Oliva and Torralba [23] data set. The authors use the original 2,688 images, initially selected for scene recognition, and added binary (yes/no) interestingness annotations via crowd-sourcing on Amazon Mechanical Turk[7]. On average, each image was annotated by 11.9 subjects. Another relevant example is the visInterest data set, Soleymani [24]. It is composed of 1,005 images covering different topics and extracted from real-world photos from Flickr[8]. Annotations were also carried out via crowd-sourcing on Amazon Mechanical Turk. Besides interestingness, these data also come with the annotation of other subjective concepts, e.g., quality, comprehensibility.

For videos, Jiang et al. [25] propose a data set consisting of 420 YouTube[9] advertisement videos extracted from 14 different categories and 1,200 Flickr videos for 15 different categories. The average duration across all the videos is around 53 seconds. Grabner et al. [26] create a webcam-based data set from publicly available webcam streams. It contains visual scenes from highways, public squares, urban scenes, etc. The data set consists of 20 different webcam sequences recorded at 1 frame/second, with 159 images each. The interestingness annotations were carried out by 46 trusted annotators and the interestingness score is assimilated to the fraction of people who marked it as interesting.

Another relevant initiative is the gifInterest data set developed by Gygli and Soleymani [27]. It addresses the prediction of GIF media interestingness and is based on the Video2GIF data set, Gygli et al. [28], and the Tumblr data set, Bakhshi et al. [29]. In total, it proposes 2,739 image sequences with an average duration of 4.25 seconds (at 11 frames/second). Annotations were computed via crowd-sourcing on Amazon Mechanical Turk.

Although existing resources are definitively valuable and address several useful use case scenarios, we propose a more comprehensive collection of resources, i.e., both annotated data and baseline systems, which were already validated in benchmark campaigns.

We identify the following main contributions over the current state of the art:

- (i) We release publicly a *consistent annotated data set*, i.e., Interestingness10k, composed of 9,831 images and 9,831 short videos (up to 4 hours), annotated for visual interestingness by trusted annotators. Apart from image and video visual interestingness, the data also allow the study of the correlation between the two. To the best of our knowledge, this is the most complete common evaluation framework available so far;
- (ii) We provide an in-depth analysis of the crucial aspects of visual interestingness prediction algorithms by investigating the capabilities and evolution of existing systems (e.g., analysis of relevant approaches from the MediaEval benchmark and from literature, influence of the employed features and fusion techniques, influence of deep learning approaches, generalization capabilities). This is again the first comprehensive study covering all these core aspects. It is a practitioners' guide for best practice in this field and also a strong baseline;
- (iii) We investigate the possibility of creating automatic, ad-hoc systems, based on existing baselines that would allow to boost state-of-the-art performance. In this context, we propose a new deep MLP-based fusion scheme that exceeds current performance by a large margin.

We analyzed the importance of this particular topic in the research community by quantifying the amount of papers published on this subject between 2010 and 2019. Results are presented in Figure 1. The search was conducted via Google Scholar[10] using the following keywords: "visual interestingness", "image interestingness", "video interestingness", "media interestingness" and "interestingness prediction". Results were filtered out to remove irrelevant articles. Although not exhaustive, it is a good approximation of the general trend. Since 2016, the first year of the interestingness prediction task at MediaEval, the number of research papers published on the subject has grown substantially, and remained high in 2019 even though the task ended. This shows the positive impact of these data, as well as the increased interest in this subject.

Relation to previous work. Some preliminary contributions of this work have been published and readers can refer to those works for more information: Demarty et al. [14,30] short papers presenting briefly the data, metrics and evaluation methodologies for the 2016

[7] https://www.mturk.com/
[8] https://www.flickr.com/
[9] https://www.youtube.com/
[10] https://scholar.google.com/

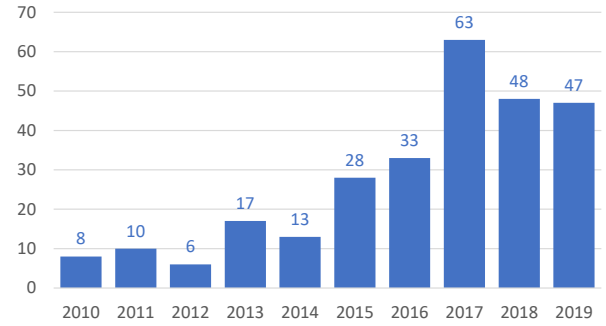

Fig. 1 Evolution of the number of published research papers referring to visual interestingness (search made via Google Scholar using "visual interestingness", "image interestingness", "video interestingness", "media interestingness" and "interestingness prediction" as keywords).

and 2017 MediaEval benchmark campaigns; Demarty et al. [31] book chapter presenting and analyzing the results of the 2016 MediaEval benchmark.

Abbreviations. Throughout the article, we employ the following abbreviations: API — application programming interface, BN — batch normalization, BTL — Bradley-Terry Luce, CNN — convolutional neural networks, C3D — convolutional 3-dimensional, CSP-RNN — circular state-passing recurrent neural network, DNN — deep neural networks, GMM — Gaussian mixture models, HMM — hidden Markov models, HoG — histograms of oriented gradients, HMP — histogram of motion patterns, HSV — hue-saturation-value, kNN — k-nearest neighbours, LBP — local binary patterns, LSTM — long short-term memory, MLP — multi-layer perceptron, MFCC — mel-frequency cepstral coefficients, mAP — Mean Average Precision, NMMP — neighborhood minmax projections, NN — neural network, PCA — principal component analysis, SIFT — scale invariant feature transform, SVM — support vector machines, SMR — supervised manifold regression, VOD — video on demand, VSEM — visual-semantic embedding model.

3 Interestingness10k data set

We present the proposed data set, its composition and annotation, inter-annotator agreement analysis and the pre-computed content descriptors provided with the data.

3.1 Composition

Interestingness10k[11] is a large-scale collection of images and video sequences extracted from Creative Commons[12] Hollywood-like movie trailers and excerpts, that allow redistribution. Trailers provide a high diversity of content with a good balance between interesting scenes and common scenes, which are typically alternating to increase the excitement. Therefore, they are more effective for generating benchmark data. Not least, having the data publicly available is a requirement for a useful benchmark. This would not be possible with data extracted from copyrighted movies.

The data are divided into two parts: (i) one for *image visual interestingness* prediction which consists of key-frames extracted from video shots[13] (middle frames), and (ii) one for *video visual interestingness* prediction which consists of individual video shots. Although images and video sequences are issued from the same data, predicting interestingness for images and videos are different tasks. Motion is characteristic for video content and affects differently the visual perception compared to a static image. Being composed this way, the data will allow the analysis of the correlation between the two. Each datum is also divided into a development set (*devset*) intended for training the methods and a test set (*testset*) for the actual evaluation. An overview of the data is presented in Table 1.

For the 2016 data, all samples are collected from 78 movie trailers. The *devset* consists of 5,054 images and 5,054 videos extracted from 52 trailers. The *testset* consists of 2,342 images and 2,342 videos extracted from 26 trailers. The 2017 data are built incrementally on the 2016 data. The *devset* data are the full 2016 data set, i.e., 7,396 images and 7,396 videos extracted from 78 trailers. The *testset* data consist of 2,435 images and 2,435 videos extracted from 26 trailers and 4 full movie excerpts. We decided to include also longer segments, e.g., the video samples extracted from the 4 movie excerpts are on average 11.4 seconds long compared to around 1-2 seconds for the others. Interestingness10k provides a total of 9,831 images and 9,831 short videos extracted from 104 trailers and 4 movie excerpts.

In Table 2, we compare our data with the most relevant data sets from literature (see also Section 2). For the image data, Interestingness10k has the advantage of providing the greatest number of images. Also, the annotations are performed by trusted annotators. For the video data, Interestingness10k provides the greatest number of sequences as well. The average duration of the video samples is slightly shorter than for the other data but it is consistent with the task. Interestingness10k is also the only data set to provide annotations for both image and video predictions.

[11] The Interestingness10k data set is available for download here: `https://www.interdigital.com/data_sets/interestingness-dataset`.

[12] `https://creativecommons.org/`

[13] A video shot is a sequence of images recorded continuously between a camera turn on and off.

Table 1 Interestingness10k basic statistics: devset stands for the development data, testset for the test data, #movies are the number of movies from which the annotated samples were extracted, and avg.dur. is the average duration of the segments. The 2016 and 2017 columns indicate the version of the data set.

Subset	*2017*	*2016*	*#movies*	*#samples*	*avg.dur.* (s)	*#interesting*	*%interesting*
Image	devset	devset	52	5,054	-	473	9.36
		testset	26	2,342	-	241	10.29
	testset		26 & 4	2192 & 243	-	261 & 55	11.91 & 22.63
Video	devset	devset	52	5,054	1.06	420	8.31
		testset	26	2,342	1.05	226	9.65
	testset		26 & 4	2192 & 243	2.15 & 11.4	249 & 28	11.35 & 11.52

Table 2 Overview of the most relevant interestingness data sets: avg.dur. is the average duration of the segments, trusted annotations are carried out by human assessors familiarized with the task.

data set	*media type*	*#samples*	*avg.dur.* (s)	*annotations*
Scene Categories - Gygli et al. 2013 [22]	images	2,688	-	crowdsourced
visInterest - Soleymani 2015 [24]	images	1,005	-	crowdsourced
Interestingness10k (proposed)	images	9,831	-	trusted
Flickr&Youtube data set - Jiang et al. 2013 [25]	videos	1,620	53	trusted
gifInterest - Gygli and Soleymani 2016 [27]	gifs	2,739	4.25	crowdsourced
Webcam - Grabner et al. 2013 [26]	sequences	20	159 (@1fps)	trusted
Interestingness10k (proposed)	videos	9,831	1.32	trusted

Initialization: assign items randomly in a matrix;
Processing:
repeat
  Perform single annotation round with multiple annotators according to the item pairs given by the square (across rows and columns);
  Compute BTL scores for the new annotations;
  Re-arrange the matrix so that items are ranked according to their BTL scores, and placed in a spiral. This arrangement ensures that similar items are compared row-wise and column-wise;
until *convergence*;

Algorithm 1: Proposed adaptive square design annotation approach.

3.2 Annotations

Annotations were performed manually by trusted human assessors, i.e., experts with good understanding of the required task. Annotations are binary, i.e., either the content is *interesting* or not. Given the fact that the image visual interestingness prediction is different than video interestingness prediction, the two annotation tasks were carried out separately.

3.2.1 Annotation protocol

We employed a *pair-wise comparison* approach, i.e., the human assessors were provided with two competing samples at a time, rather than annotating individual items, a method well suited for gathering subjective annotations in similar scenarios, as presented by Salesses et al. [32]. This provides several advantages. Firstly, it is more reliable as the annotator is asked to do a relatively easy cognitive task, i.e., simply comparing two items. In theory, assigning an absolute rating for a single item requires the annotator to compare to the full set of previously seen items, or at least to keep in mind some complicated set of decisions, Yang and Chen [33]. Secondly, for independent items, different annotators may use different scales and the assessments are not easily comparable, Ovadia [34]. Finally, it has been shown that pairwise comparisons are less influenced by the order in which the annotations are displayed compared to a direct rating, Yannakakis and Hallam [35]. To comply with the underlying use case scenario, annotators were instructed to select *the image/video that would be defining for making him watch the entire source movie.*

The main drawback of a pair-wise comparison approach is the impossibility of exploring all possible combinations of two items, especially when dealing with such a large data set. There are however several approximations possible which converge to similar results. We started from the adaptive square design method, Barkowsky and Callet [36], where the items are placed in a square and only pairs on the same row or column are compared. This reduces the number of comparisons from $n(n-1)/2$ for all pairs, to $n(\sqrt{n}-1)$, where n is the number of items. The Bradley-Terry-Luce (BTL) model [37] was used to convert the paired comparison data to a scalar value. We modified the original adaptive square design setup so that comparisons were made by many users simultaneously until all the required pairs had been annotated. The proposed algorithm[14] is depicted in Algorithm 1.

[14] The web-based pair-wise annotation software tool is available here: `https://github.com/mvsjober/pair-annotate`.

For the annotations, we used 5 rounds which proved to be sufficient to achieve good convergence. The final interestingness decision was based on sorting the BTL values and finding a threshold value. We used a heuristic rule to find the boundary between the interesting and non-interesting items, i.e., normalizing the BTL values for each movie separately and using the assumption that the BTL distribution is a sum of interesting and non-interesting sample distributions. For more details about the protocol, see Demarty et al. [31].

3.2.2 Annotation statistics

The image data set was annotated by 270 annotators (average age 25.2±9) for which 70.9% were males, and 29.1% were females. Annotators came from 17 different countries around the world, mainly from Europe (79.6%) and Asia (18.5%). On average, each annotator annotated 1,976 different image pairs. The video data set was annotated by 526 annotators (average age 30.3±12.5). The gender distribution was similar to the one for the image data, with 66.7% males and 33.3% females. Annotators were spread over 35 countries, distributed slightly different compared to the image annotators, namely 74.5% came from Europe, 15% from Asia, 8.7% from America and 1.7% from the rest of the world. On average, each annotator annotated 1,030 video pairs, which is approximately half the number of image pairs. The reason is the significantly longer time required to visualize the videos.

Given the high subjectivity of the task, it is interesting to assess the annotators' agreement. To do so, there is a high diversity of metrics available, e.g., Percent Agreement, Krippendorff's alpha, Fleiss Kappa, Randolph's kappa, Hayes and Krippendorff [38]. Depending on the data type and size, their characteristics, the number of raters per sample, not all metrics are suitable and equivalent. In our case, we have a large collection of annotations with 533,520 pair annotations for images and 541,780 pair annotations for videos. Not all pairs were viewed by the same annotators, but all of them had votes from at least two different annotators. Inter-rater agreement's measures such as Fleiss' kappa or Randolph's kappa are particularly appropriate in such configuration. Furthermore, the annotations are not equally spread between the two categories, i.e., *interesting* and *not interesting*. We observed a bias towards the *not interesting* class for both images and videos, with only a few samples with high interestingness levels. In the adopted pair-wise comparison protocol, there were no constraints adopted to attempt to equally spread the data into the two classes.

In such cases, where raters don't know *a priori* the number of cases that should be distributed into each category, Randolph's kappa proved to be a good alternative to the fixed-marginal multirater Fleiss' kappa, Randolph [39]. Marginals are considered to be fixed when raters know *a priori* the quantity of samples that should be distributed into each class. In that sense, Randolph's kappa is seen as a free-marginal multirater kappa, adapted to a non-symmetric distribution of the data between classes.

The computation of Randolph's kappa, when considering two annotators per pair, led to a value of 0.556 for the image data set and 0.519 for the video data set. Randolph's kappa is in the range of $[-1; 1]$, with 1 being a perfect agreement and negative values meaning no agreement between raters (other than what would be expected by chance). Therefore, we reach a reasonable agreement on both the image and the video data sets. For the sake of comparison, we also computed the Percent Agreement and obtained 76.9% for the image data set and 75% for the video data set. This reconfirms a reasonable inter-rater agreement for both data sets, considering the high subjectivity of the interestingness concept.

In Figures 2 and 3, we illustrate several examples of both images and videos, annotated as interesting as well as non-interesting. Interesting content is visibly more colored, better centered on pleasant people, less blurred and containing interesting actions.

3.3 Content descriptors

To address a broader community, the data come with several pre-computed, general purpose, content descriptors for visual and audio information.

Visual information. We propose the following visual descriptors: Dense SIFT, Lowe [40], HoG, Dalal and Triggs [41], LBP, Ojala et al. [42], GIST, Olivia and Torralba [23], Color Histogram, AlexNet layers, Krizhevsky et al. [43], and C3D layers, Tran et al. [44]. Dense SIFT features were computed using densely sampled frame patches instead of point of interest detectors, with a codebook of 300 codewords used in the quantization process, as described in Lazebnik et al. [45]. HoG descriptors were computed over densely sampled patches and following the work of Xiao et al. [46] were concatenated in order to create a higher dimensional feature. From the AlexNet model we extracted the fc7 and prob layers, according to the work of Jiang et al. [47] and from the C3D model the fc6 layer.

Audio information. We propose the MFCC features, computed over 32ms windows with 50% overlap, where

Fig. 2 Examples from Interestingnes10k image data set: images annotated as interesting are on the right, whereas non-interesting images are on the left (source: videos 1 to 7). Images at the top have higher annotator agreement, while images at the bottom have lower annotator agreement.

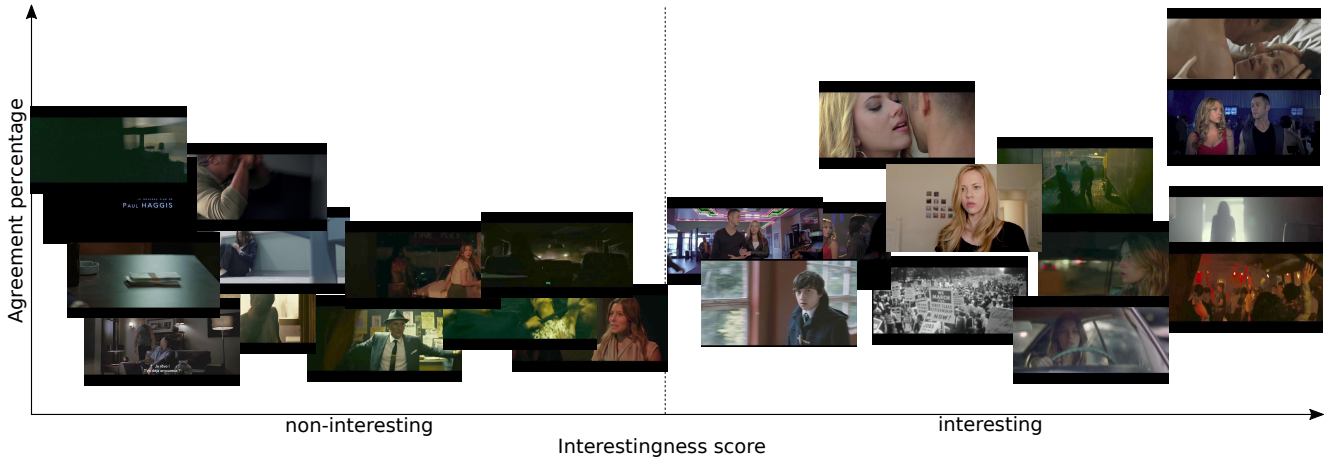

Fig. 3 Examples from Interestingnes10k video data set: videos annotated as interesting are on the right, whereas non-interesting videos are on the left (source: videos 1 to 7). Videos at the top have higher annotator agreement, while videos at the bottom have lower annotator agreement. Each video is depicted with a key-frame.

cepstral vectors are concatenated with their first and second derivatives.

Mid-level information. To account for a higher-level description, we propose a human presence detector. Face detection was computed via HoG and tracking was done via the approach proposed by Danelljan et al. [48].

4 Evaluation methodology

For benchmarking image/video visual interestingness prediction, we recommend certain metrics. These were used during the 2016 and 2017 MediaEval benchmark. Of course, the data are not restricted to those ones, but they provide a solid baseline. There is also an official split between the training data (*devset*) and testing data (*testset*). It is presented in Table 1. This would allow systems to be compared under the same conditions. The systems should train their parameters on *devset* and perform the actual evaluation on the *testset*.

We expect the systems to predict a confidence score corresponding to the degree of visual interestingness for each item. The higher the score, the more interesting it should be. Inline with this, we recommend two related metrics: the overall mean Average Precision (mAP) and the mean Average Precision over the 10 highest ranked items (mAP@10). MAP is a widely used metric for retrieval tasks, proven to be stable in such scenarios, Buckley and Voorhees [49]. It is computed as the mean value over the average precision scores for each source trailer in the *testset*.

This metric fits the VOD use case where images/videos should be selected to be the most interesting for representing the underlying content. mAP@10 was proposed to better reflect the selection of a small set

Table 3 Number of systems analyzed (MediaEval refers to systems submitted to the benchmark whereas state-of-the-art refers to relevant systems from literature).

Data set	*MediaEval*	*State-of-the-art*
2016.Image	27	12
2016.Video	27	18
2017.Image	33	5
2017.Video	42	28

of candidate images/videos. The metrics are computed using the standard treceval software tool[15].

5 Baseline systems

We provide an in-depth analysis of various systems, both from the MediaEval benchmark, as well as state-of-the-art systems from literature which were evaluated on Interestingness10k. We reference the different year data as $Year.Type$, where $Type$ is the modality and $Year$ the specific year, e.g., 2016.Image refers to the 2016 image prediction data. The data were presented in Table 1. We overview a total of 192 systems, as presented in Table 3.

Systems are evaluated using the official *devset-testset* split and also the official metrics, i.e., mAP for the 2016 data and mAP@10 for the 2017 data. For comparison between different data sets, we use general mAP.

We provide an analysis of overall performance and system evolution, employed descriptors, prediction methods, generalization capabilities and, finally, analyze the reliability of the system ranking results. Statistical significance of the main hypotheses are tested using the Mann-Whitney-U test [50].

5.1 Analysis of the overall performance

We analyze the general trends and performance of existing approaches. A boxplot representation of the results is presented in Figure 4.

The first observation is the fact that no methods stood out as outliers, i.e., with significantly higher or lower performance, compared to the others. There is then an obvious trend of increasing performance from 2016 to 2017. The best mAP performance for image prediction increases by 25.75%, from 0.2485 on 2016.Image data, Constantin and Ionescu [51], to 0.3125 on 2017.Image data, Parekh et al. [52].

For video prediction, the improvement is similar, namely 22.75%, from a mAP value of 0.1815 on 2016.Video data, Almeida [53], to 0.2228 on 2017.Video data,

[15] `https://trec.nist.gov/trec_eval/`

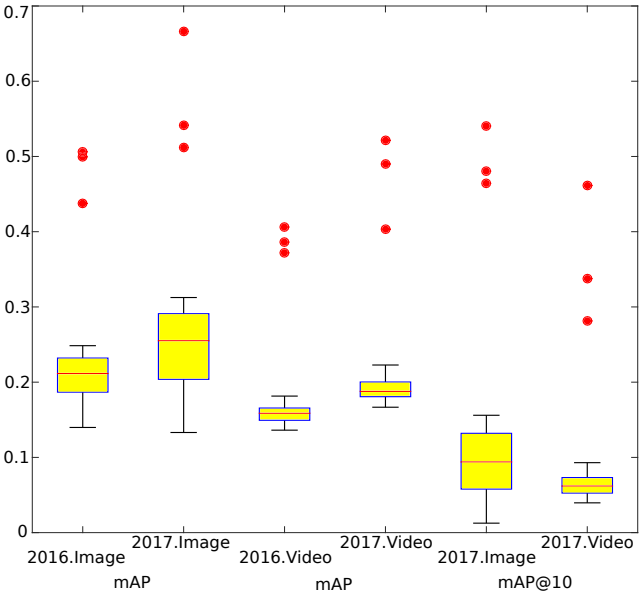

Fig. 4 Boxplot representation of the overall performance: interquartile range (IQR) 50%, median values (red line), lower and upper adjacent values calculated as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ respectively. For reference, the performance of 3 human runs is represented with the red dots.

Wang et al. [54]. The median mAP for 2017.Image and 2017.Video, 0.2550 and 0.1877, respectively, both surpass the maximum values recorded for 2016.Image and 2016.Video. The observation is also true when analyzing only the runs that were officially submitted as part of the MediaEval benchmark, with the image and video prediction registering a growth of 31.63% and 15.37%, respectively (Mann-Whitney-U $p < 0.001$ for images and videos). The reason behind this could be the improvement of the systems, their increased specialization for an interestingness related task, the effect of a bigger number of samples in the training data sets and better annotations.

Comparing the prediction of visual interestingness for images and videos, results show that images allow to achieve higher mAP, but also a wider spread of the results meaning more diversity, i.e., standard deviation is 0.0264 for Image.2016 data vs. 0.0120 for Video.2016 data, and 0.0476 for Image.2017 data vs. 0.0134 for Video.2017 data.

To stress the upper limit performance, we also assess the results of three human runs, obtained via the human annotators (see the red dots in Figure 4). To compute those, we followed the annotation protocol described in Section 3.2.1. The best achieved results are: for the 2016.Image data a mAP of 0.5058, for the 2016.Video data a mAP of 0.4066, for the 2017.Image data a mAP@10 of 0.5403 and a mAP of 0.6661, and for the 2017.Video data a mAP@10 of 0.4140 and a mAP

of 0.4897. What is interesting to notice is that these human assessors did not lead to 100% precision, as the overall aggregated annotations would do. This clearly indicates the high subjectivity of such a task and inherently a variation in the perception of the data.

5.2 Analysis of the employed features

We further analyze the impact of the employed content descriptors and fusion schemes on the performance.

5.2.1 Per feature type analysis

We identified 6 prominent modalities which are exploited, namely: *visual* (e.g., HoG), *audio* (e.g., MFCC), *motion* (e.g., HMP,), *deep features* (e.g., AlexNet layers), *conceptual* (e.g., SentiBank) and *text* (e.g., movie metadata). We selected 18 of the employed combinations which provided the most interesting results. They are presented in Figure 5.

Single modality features. Overall, 72% of the analyzed systems (139 systems) use only one modality. Some have achieved very good performance. For instance, Constantin and Ionescu [51] achieve the best overall performance on the 2016.Image data set, with mAP 0.2485. The authors use a combination of standard *visual features* (Datta et al. [55], Li and Chen [56], Ke et al. [57]) with early and late fusion schemes. *Motion features* were best overall performers on the 2016.Video data. Almeida [53,58] achieves a mAP of 0.1815 using histograms of motion patterns [59] in different learning-to-rank strategies, thus taking into account the full spatio-temporal representation of the videos.

High level concepts. Conceptual features are a special class of descriptors that represent higher level concepts, positively or negatively correlated with interestingness. Even though few systems have implemented such features, only 12% (23 systems), they achieve some of the top results. Examples are features capturing emotions, e.g., SentiBank, Borth et al. [60] employed by Xu et al. [61], features representing the visual-semantic space, e.g., image-captioning based, Kiros et al. [62], employed by Berson et al. [63]. The best results on 2016.Image data from the MediaEval benchmark was achieved by Liem [64] who uses HSV *histograms* augmented with the presence and areas of *faces*. It achieves a mAP of 0.2336. On the 2017.Video data the best mAP@10 is 0.0827, achieved by Ben-Ahmed et al. [65] at MediaEval. The authors use *genre* as a predictor for movie interestingness, developing a system that creates genre predictors based on layers extracted from deep neural networks like VGG-16, Simonyan and Zisserman [66], and SoundNet, Aytar et al. [67]. The proposed system uses the MovieScope data set, Sivaraman and Somappa [68] as additional training information.

Deep features. Deep features are now the state of the art in many classification tasks. They were also widely used, both as unimodal features or part of multimodal, fusion approaches, accounting for 59% of the analyzed systems (114 systems). Examples are the use of AlexNet fc7 and prob layers in Erdogan et al. [69], or last layers of VGG in Lam et al. [70]. Overall, several deep feature-based systems achieved the best performance, either individually or in multimodal combinations. The highest mAP on 2016.Image data achieved during the MediaEval benchmark, is 0.2336 and is obtained by Shen et al. [71]. The authors employed *fc7 layer features* from CaffeNet [72], where data are re-sized and center cropped to preserve the aspect ratio. The authors also performed a mean image subtraction for normalization. Another example is the approach of Parekh et al. [52], which achieved the best overall mAP@10 on 2017.Image data, i.e., 0.156. The authors use the *fc7 layer* of AlexNet as input for their DNN ranking. Other approaches use deep features in fusion schemes. The best mAP@10 on 2017.Image data achieved during the MediaEval benchmark is 0.1385. Permadi et al. [73] employed standard *visual* features like LBP and HoG in combination with AlexNet *fc7 features*. The best overall result on the 2017.Video data, mAP@10 0.093, is achieved by Wang et al. [54] fusing standard *visual* features (color histogram, denseSIFT, GIST, HoG and LBP), *audio* features (IS10, Eyben et al. [74]) and *layers* of deep networks (AlexNet, C3D and InceptionV3).

Feature aggregation. A type of feature fusion was employed by 54% of the analyzed systems (104 systems). Early fusion was used in 41% of the cases (78 systems), while late fusion was used in 25% of the cases (48 systems). Some systems use a combination of these approaches. Dimensionality reduction schemes have been used in 18% of the cases (34 systems). Some notable performance is achieved including PCA, in Rayatdoost and Soleymani [75], with the third best result on 2016.Image data at MediaEval (mAP 0.1710), and NMMP and SMR, in Liu et al. [76], with the second best mAP result in 2017.Image at MediaEval(mAP@10 0.1369). For a detailed analysis of the impact of *dimensionality reduction* on the prediction, the reader can refer to Liu et al. [77].

Temporal feature aggregation was also explored. Several methods were tested for creating video level descriptors from individual frame descriptors. Overall, 67 out of the total 115 systems (58%) dealing with video interestingness use this type of feature aggregation. How-

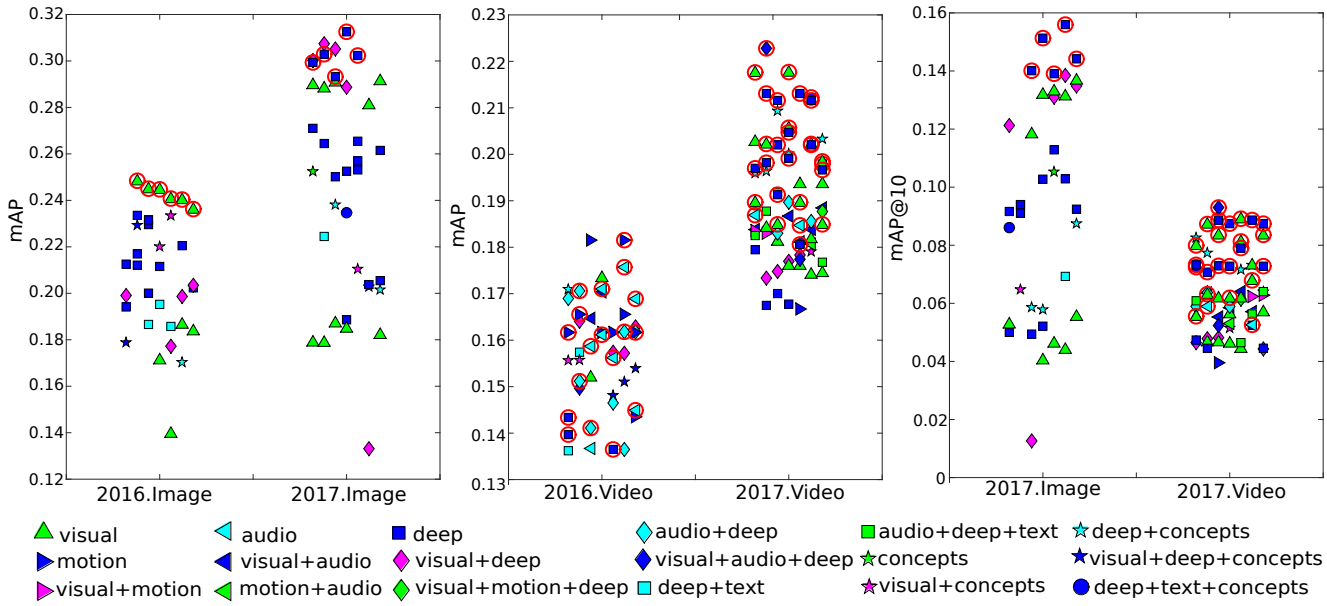

Fig. 5 Analysis of the employed features: Year.Type represents the year of the data (2016 or 2017) and its type (Image or Video). Official metrics for 2016 data is mAP and for 2017 is mAP@10. For comparison, we also provide mAP for 2017. We represent both the participating systems from MediaEval benchmark as well as state-of-the-art approaches from literature (marked with a red circle).

Table 4 Average mAP over the analyzed systems for each feature category: visual, deep, motion, audio, text, and concepts, and fusion scheme: no fusion, early fusion, and late fusion.

		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

ever, most of them were traditional *statistical* methods, such as average and standard deviation. For instance, Liu et al. [76] obtains the second best mAP during MediaEval for 2016.Video with a mAP of 0.1735. Median is used in Constantin et al. [78], obtaining a mAP@10 of 0.0732 on 2017.Video data. This is the third best run at MediaEval. There are also some interesting approaches like the use of *Bag-of-Features* in Almeida and Savii [79], who achieved a mAP@10 of 0.0628 on 2017.Video data, or the use of *temporal integration* via LSTM [80] architectures in Shen et al. [81], with a mAP of 0.1706 on 2016.Video data.

5.2.2 Overall feature performance analysis

Table 4 presents an analysis of the average mAP achieved with the six modalities presented in the previous section. For the image data, systems that have incorporated deep features perform better on average, with an average mAP of 0.2297, while for the video data, systems that use traditional visual features perform better, with an average mAP of 0.1798. An interesting trend can be observed when looking at the fusion schemes employed by these methods. Late fusion approaches tend to perform better than others, with an average mAP of 0.2416 for the image data and of 0.1878 for the video data. This supports the idea of employing more advanced late fusion techniques to significantly boost the performance (see Section 7).

5.3 Analysis of the prediction methods

The next experiment is to analyze the employed techniques and their capabilities. There are, of course, numerous approaches that have been experimented. However, we can identify some trends. We propose an analysis at two different levels of detail, the methods being classified: (i) according to the problem formulation, and (ii) according to the specific class of techniques. While some of the classes defined in the following section may not be mutually exclusive, our intention here is not their classification but to identify the most prominent approaches and understand their performance and general trends.

Fig. 6 Analysis of the employed methods: Year.Type represents the year of the data (2016 or 2017) and their type (Image or Video). We plot mAP for all methods. We provide two levels of details: (i) per problem formulation, and (ii) per specific method. We represent both the participating systems from MediaEval benchmark as well as state-of-the-art approaches from literature (marked with a red circle).

5.3.1 Problem formulation

We identify the following main approaches: (i) *classification*, (ii) *regression*, (iii) *ranking*, and (iv) *hybrid*, i.e., combining more than one approach. Results are presented in Figure 6.

Overall, more than 52% of the approaches use *classification*. These systems tend to achieve the highest performance on the image data. For example, Constantin et al. [78] and Shen et al. [71] achieve a mAP of 0.2485 and 0.2336, respectively, on the 2016 data. *Ranking* approaches account for about 31% of the employed systems. They are used by the two top runs on the 2016.Video data, i.e., Almeida et al. [53,58] with a mAP of 0.1815. It is worth noting that ranking approaches achieve overall better results in all the cases compared to the classification (Mann-Whitney-U $p < 0.005$). *Regression* approaches were used by 15% of the total number of analyzed systems. They were used in one of the top runs on the 2017.Image data. Permadi et al. [73] achieves a mAP@10 of 0.1385 at MediaEval. *Hybrid* approaches account for about 2% of the total number of systems. Despite being the less popular ones, they achieved the best overall results on the 2017.Video data, e.g., Wang et al. [54] with a mAP@10 of 0.093.

5.3.2 Approach

Several general trends can be observed while analyzing the methods according to their specific techniques. Of course, different classifications can be made. We propose the following categories which are consistent with the results: (i) *Support Vector Machines*, (ii) *Deep Neural Networks* (e.g., VSEM, Video2GIF, LSTM-ResNet, CSP-RNN), (iii) *ranking* (e.g., RankNet, RankBoost), (iv) *regression* (e.g., Support Vector Regressor, Logistic Regression, Supervised Manifold Regression), (v) *hybrid* approaches combining more than one method (e.g., Nearest Neighbor and Support Vector Regressor, SVM and Ranking-SVM, Siamese network and Markov Decision Process), (vi) *Neural Networks* (e.g., Multi-Layer Perceptron), (vii) *distance-based* approaches (e.g., Nearest Neighbours), (viii) *ensemble learning* (e.g., Random

Forest), and (ix) *statistical* approaches (e.g., Markov Decision Process). The results are presented in Figure 6.

Support Vector Machines. SVM was the most popular choice among the analyzed approaches, representing 30%. It is used by two of the top runs on the 2016.Image data, and the top run on the 2017.Video data. For instance, Shen et al. [71] built a simple yet efficient system that uses visual features (CNN features from the last layer of the CaffeNet) classified with SVMs. It achieves a mAP of 0.2336 on the 2016.Image data during the MediaEval competition. This score was further improved by Constantin and Ionescu [51] using SVMs to learn the association between various image description techniques (related to subjective properties, such as aesthetics, style, image composition) and interestingness. The system is boosted via a late fusion approach, outperforming the best results from the MediaEval benchmark with a mAP of 0.2485. On the 2017.Video data, the top score at MediaEval is obtained by Ben-Ahmed et al. [65] using deep audio-visual features to generate mid-level concepts representing movies genres, i.e., action, drama, horror romance, and sci-fi. These genre distributions served as the input for a binary SVM classifier achieving a mAP@10 of 0.0827.

Deep Neural Networks. DNNs represent 28% of the number of analyzed approaches, being the second most used approach. Surprisingly, none of the best MediaEval benchmark systems are using DNNs. However, there are state-of-the-art approaches which outperform the best results. For instance, Parekh et al. [52] provides the best overall results on 2017.Image data, achieving a mAP@10 of 0.156. The authors train a DNN network that takes as input pairs of CNN representations of images, to predict which one is more interesting from the pair. The process is carried out for all possible pairs within each video followed by a ranking algorithm.

Ranking. Ranking approaches account for 13% of the analyzed approaches. Almeida et al. [53] uses a set of learning-to-rank algorithms for predicting the interestingness of videos via only visual feature representations (HMP). The classification is performed with a majority voting scheme over the prediction of 4 pairwise learned rankers, namely: Ranking SVM, RankNet, RankBoost, and ListNet. It achieves the best results in the MediaEval competition on the 2016.Video data, i.e., mAP 0.1815.

Regression. Regression approaches account for 12% of the analyzed systems, while also accounting for some top runs. For instance, Permadi et al. [73] achieves the best overall results on the 2017.Image data, a mAP@10 of 0.1385. The authors use a logistic regression trained on an early fusion representation of various features, i.e., Color Histogram (HSV), LBP, HoG, GIST, denseSIFT, Alexnet features and contextual descriptors.

Hybrid. Hybrid approaches, combining more than one type of methods, account for almost 6% of the total analyzed systems. While these methods did not achieve notable results during the MediaEval benchmark, some of the state-of-the-art approaches provide notable results. Wang et al. [54] provide the best overall results on 2017.Video data, with a mAP@10 of 0.093. The authors investigate the use of a learning-to-rank DNN via a Siamese network, and a reinforcement ranking based on a Markov decision process. To boost the results, descriptors are aggregated using early fusion: visual descriptors (GIST, LBP, HoG, Color Histogram, denseSIFT), deep features (AlexNet, InceptionV3, C3D), and acoustic features (energy, pitch, jitter and shimmer). A late fusion is finally used to aggregate the decisions of the two ranking models.

Neural Networks. Shallow NN-based methods are less used and account for almost 6% of the analyzed systems. While in general less effective than the other approaches, one approach stood out. Berson et al. [63] uses semantic and contextual information via CNN features and image-captioning based features with metadata extracted from IMDb[16]. The authors investigate different combinations of features trained via a simple MLP network, achieving a mAP@10 of 0.1054 on the 2017.Image data.

Distance-based. Distance-based approaches account for 4% of the total number of analyzed systems. For instance, Liem et al [64] employs a heuristic approach based on the occurrence of people in video shots. The author assumption is that clear human faces should attract viewers attention causing larger empathy. The classification quantifies the average of the histogram intersection between the HSV histograms of the detected faces, the mean HSV of all frames with detected faces within a shot, and the area of the detected faces' bounding boxes. The scores are then sorted followed by thresholding. It achieves a mAP of 0.2336 on the 2016.Image data and 0.1558 on the 2016.Video data.

Ensemble learning. Ensemble learning approaches are poorly represented. We find only one approach tested on the 2016.Image data but without any notable results.

Statistical. Similarly, statistical approaches, e.g., Markov decision based, were used by only one system on the 2017.Video data, but without any notable results.

5.3.3 Overall method performance analysis

Table 5 presents an analysis of the average mAP achieved for the categories of methods presented in the previ-

[16] `https://www.imdb.com/`

Table 5 Average mAP over the analyzed systems for each category of methods: Support Vector Machines (SVM), Deep Neural Networks (DNN), ranking, regression, hybrid, neural networks (NN), distance-based approaches (distance), ensemble learning (ensemble), statistical approaches.

		SVM	DNN	ranking	regression	hybrid	NN	distance	ensemble	statistical
Image	avg. mAP	0.2269	0.2460	0.2374	0.2242	0.1854	0.2405	0.2214	0.2011	-
	#systems	27	13	6	12	2	6	3	2	0
Video	avg. mAP	0.1822	0.1799	0.1712	0.1666	0.1867	0.1848	0.1585	-	0.2021
	#systems	28	39	18	10	9	5	4	0	2

ous sections. For the image data, approaches based on DNNs and shallow NN stand out, with average mAP scores of 0.2460 and 0.2405, respectively. This result is particularly interesting as the best performing type of method, DNN, has also a high number of runs. While the most used approach is SVM, it is outperformed by many of the other approaches. On the other hand, for the video data, hybrid approaches and SVM-based approaches stand out as the best performers, with average mAP scores of 0.1867 and 0.1822, respectively. Unlike the image data, it appears that hybrid systems are the best performing type of methods, which could be the result of the inherently multi-modal nature of videos.

5.4 Generalization capabilities

Interestingness has been proved to be either positively or negatively correlated to other subjective concepts, Constantin et al. [4]. It is therefore interesting to study whether systems are able to generalize well from other concepts or data, and even between images and videos. In this experiment we analyze these aspects.

5.4.1 Concept generalization

We analyze how visual interestingness prediction generalizes between different concepts and, therefore, type of data. We identified the following situations: (i) *no generalization*, i.e., the systems were trained solely on the Interestingness10k data, without the use of other external data; (ii) *pre-trained extractors*, i.e., systems are trained on data unrelated to interestingness, like object recognition data sets, and used directly, usually as features in a classifier, to predict interestingness; (iii) *fine-tuned systems*, i.e., systems are firstly trained on data unrelated to interestingness and then retrained on the Interestingness10k data to predict visual interestingness; (iv) *correlated systems*, i.e., systems are trained on other data from positively or negatively correlated domains, e.g., memorability, aesthetics, emotion prediction, and then used to predict interestingness, either directly or via finetuning.

Pre-trained extractors, with 88 systems (45.8%), represent the most common type of system, even more popular than systems that do not use any kind of generalization (42.2%). Several deep neural network architectures were used by these extractors, including AlexNet, VGG and C3D.

Fine-tuned systems were mainly employed by finetuning popular deep neural networks, accounting for 17 systems in total (8.9%), 8 of them addressing image interestingness and 9 of them video interestingness. For instance, Erdogan et al. [69] achieves a mAP of 0.2125 on the 2016.Image data. The authors fine-tune the AlexNet model. The last softmax layer is replaced with a regression layer, using Euclidean loss. Training is carried out for 2,000 epochs and only the weights of the final fully connected layer are updated during this process. Ben-Ahmed et al. [65] achieves the best results on 2017.Video data with a mAP of 0.2094, being also the best result recorded during the MediaEval benchmark. The authors create a genre prediction system for video and audio information using the VGG and SoundNet models, trained on the MovieScope data set [68]. The final retrained system is able to infer video interestingness starting from the genre prediction network. During the training process, keyframes were used as representatives for the entire video shot. Another approach, developed by Vasudevan et al. [82], uses a deep visual semantic embedding model developed and trained on 0.5 million samples from the MSR Clickture data set [83], used to infer semantic proximity between text and images. This network uses a series of LSTM layers for encoding textual information and convolutional and fully connected layers for image processing. During the finetuning process, the title of the movie and the keyframes are embedded in the same space, and ranking is achieved based on the distance between the textual and image embeddings. This approach scored a mAP of 0.1952 on the 2016.Image data.

Finally, 6 systems (3.1%), 3 for image prediction and 3 for video prediction use *correlated system* approaches. For image prediction, Shen et al. [71] achieve a mAP of 0.2315 on the 2016.Image data. The authors create a shallow MLP-based system with one hidden dense layer with 1,000 neurons and ReLU activation. This system is initially trained on a data set of 0.2 mil-

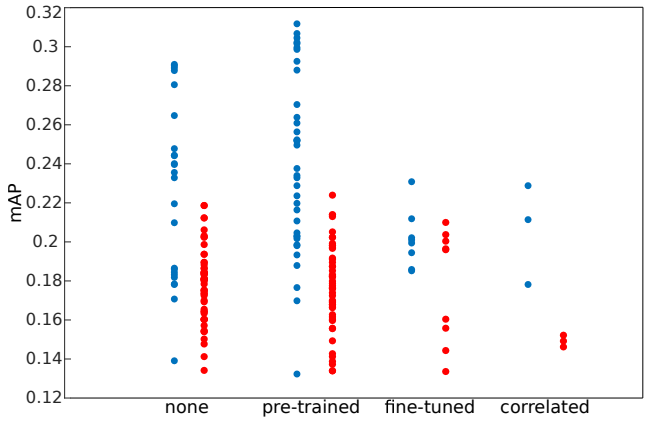

Fig. 7 Analysis of the generalization capabilities: methods developed on the provided data (*none*), methods pre-trained on unrelated data (*pre-trained*), methods pre-trained and then re-trained on provided data (*fine-tuned*), methods pre-trained on related data and used directly (*correlated*). mAP values are presented for all the methods, whereas image prediction is depicted in blue while video prediction in red. We represent both, MediaEval benchmark systems as well as state-of-the-art approaches from literature.

lion images extracted with the Flick API[17], based on their Flickr social interestingness score. The data set is evenly balanced with regards to socially interesting and non-interesting samples. While social interestingness and visual interestingness are different concepts, they can exhibit some degree of correlation given their subjective nature [13]. The best performing model is trained afterwards on the 2016.Image data and some additional resampling and upsampling steps are taken to induce class balance.

For the upsampling strategy, interesting samples are multiplicated, by a factor of 5 to 13 times, with the optimum result being achieved for an upsampling factor of 11. This approach also represents the best result attained during the MediaEval competition, with a mAP of 0.2336. For the resampling strategy, the authors randomly select samples, based on a preset probability of interesting samples being selected. Values between 0.3 and 0.6 interesting samples are tested, and the optimum result, a mAP of 0.2315 is achieved with a resampling parameter of 0.6. Other approaches include the ones proposed by Erdogan et al. [69], who retrain the fully connected weights of the memorability model MemNet [84] for 3,000 epochs, thus achieving a mAP result of 0.2121. For both image and video prediction, Xu et al. [61] employed SentiBank-based systems in their approach, trained on Flickr images [60], without finetuning the systems on Interestingness10k data. For the 2016.Image data, the authors achieve a mAP of 0.229, while for the 2016.Video data the result is 0.154. Previous works have shown positive correlation between emotional content and visual interestingness [22].

Figure 7 shows a comparison between the results obtained by different generalization strategies. It is interesting to notice that, for image interestingness prediction, the pre-trained extractor systems are performing significantly better than the other type of methods. The average mAP for pre-trained systems is 0.2405, while for the no generalization systems the average mAP is 0.2208 (Mann-Whitney-U $p < 0.05$). However the same conclusion did not present statistical significance for the video data. While the other strategies did not present top results, an interesting experiment is conducted by Vasudevan et al. [82]. As mentioned before, their network, once re-trained on 2016.Image data achieves a mAP value of 0.1952. However, the same deep visual semantic embedding system trained only on the 0.5 million text-image pairs only achieves a mAP of 0.1866, while the addition of 7.5 million text-image pairs from the MSR Clickture data set surprisingly further decreases the mAP to 0.1858. This experiment shows the importance of finetuning on Interestingness10k data and the performance advantage it can bring.

5.4.2 Image to video generalization

We analyze how image visual interestingness prediction can generalize to video prediction. We target identical systems, e.g., use of the same set of features, preprocessing, training and post-processing, that are used for both tasks. This analysis also incorporates video systems that use simple statistical approaches in creating a video descriptor out of image descriptors, such as taking average or median values across the entire set of frames and generating a single, video-wise descriptor. 10 systems fall into this category. Figure 8 presents the achieved mAP on video prediction vs. image prediction. The Pearson correlation coefficient is $\rho = 0.546$ indicating that there is correlation between the two. However, this can be explained also by the data which is also correlated, i.e., images are extracted from the videos.

Nevertheless, although not a statistical proof, we don't rule out the possibility of adapting image-to-video prediction and vice-versa. This was also experimented in some previous work, e.g., Liu et al. [85], where systems are adapted to both tasks.

5.4.3 Long vs. short videos

The 2017.Video data include some longer than the average videos (see Section 3.1), with an average duration of 11.4 seconds compared to around 1-2 seconds for the

[17] `https://www.flickr.com/services/api/flickr.interestingness.getList.html`

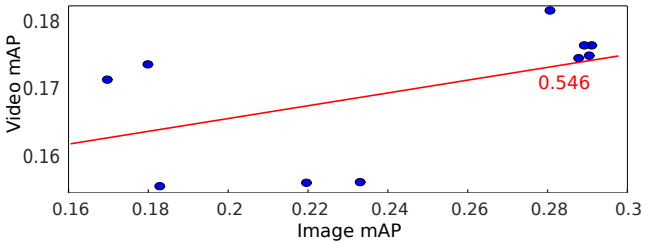

Fig. 8 Analysis of image-video generalization capabilities: mAP achieved on video prediction vs. mAP achieved on image prediction via the same approach (blue dots). Pearson correlation is 0.546 (depicted in red).

others. We analyze here the prediction capabilities between these different length data. Results prove that the longer the videos, the better the prediction of the system. The average mAP@10 on the 1-2 second videos is 0.0562, while the average mAP@10 for the 11.4 second videos is 0.0751.

5.5 Reliability analysis

We analyze the reliability of the MediaEval benchmark rankings for the Interestingness10k data. The general idea is to study how stable the rankings are by sampling the testing data set in different ways.

Systems are ranked using an evaluation metric based on comparing their responses to the ground truth for a set of queries $q \in \mathcal{Q}$. If we denote the score achieved by system A with $\lambda_{\mathcal{Q},\mathsf{A}}$, and the score received by a different system B with $\lambda_{\mathcal{Q},\mathsf{B}}$, we say that system A is better than system B if $\lambda_{\mathcal{Q},\mathsf{A}} > \lambda_{\mathcal{Q},\mathsf{B}}$. If this ranking is reliable, it could be replicated with another set of queries $\mathcal{Q}'$, so that $\lambda_{\mathcal{Q}',\mathsf{A}} > \lambda_{\mathcal{Q}',\mathsf{B}}$ still holds, Urbano et al. [86].

Ranking stability was investigated by randomly sampling equally sized pairs $(\mathcal{Q}', \mathcal{Q}'')$ of query subsets from all *testset* queries $\mathcal{Q}$. Next, the system rankings based on $\mathcal{Q}'$ can be compared with those based on $\mathcal{Q}''$. Urbano et al. [86] suggests several reliability indicators for performing this comparison, and show that most of them are highly correlated. We selected two measures for this study, representative of two different types of measures: relative sensitivity (score-based) and Kendall's rank correlation (rank-based). In addition, we also calculate a weighted variant of Kendall's rank correlation.

Relative sensitivity δ_r is defined as the minimum difference $(\lambda_{\mathcal{Q}',\mathsf{A}} - \lambda_{\mathcal{Q}',\mathsf{B}}) / \max(\lambda_{\mathcal{Q}',\mathsf{A}}, \lambda_{\mathcal{Q}',\mathsf{B}})$ that needs to be observed with $\mathcal{Q}'$ such that the differences with $\mathcal{Q}''$ have the same sign at least 95 % of the time. For a stable system, relative sensitivity tends to 0, and Sanderson et al. [87] suggest $\delta_r = 0.25$ as a reasonable limit for judging reliability.

In contrast, Kendall's rank correlation τ considers only the systems' ranks and not their specific scores, Abdi [88]. Instead, it depends only on the number of inversions of pairs of objects that would be needed to transform the ranking induced by $\mathcal{Q}'$ to the one by $\mathcal{Q}''$. The value of τ ranges from 1 (identical rankings) to -1 (inverse ranking). Voorhees [89] suggests $\tau = 0.9$ as a reasonable limit for judging the ranking reliable.

Finally, we also compute the weighted Kendall's rank correlation τ_w, Vigna [90]. Here, exchanges of highly ranked objects are considered more influential than exchanges of low ranked objects. We consider that this is well-motivated in this case as the worst systems are performing essentially randomly, and their ranking can thus be deemed somewhat arbitrary. We used the additive hyperbolic weighting as suggested by Vigna [90].

The Interestingness10k *testset* data contains around 2,400 video shots, which are extracted from 26 videos for 2016 and 30 for 2017. In order to have statistically independent subsamples we opted to sample among the set of videos, as shots from the same video cannot be considered to be statistically independent. We have subsampled in decrements of one, so that if the total number of videos is N, we have proceeded to randomly generate pairs of $N-1$ movies, $N-2$, and so on. For each subsample size we report average scores calculated across 50 randomly generated pairs.

Figure 9 shows the reliability scores for each datum and modality according to the official metric. In all plots, the horizontal axis indicates the subsampling percentage, while the vertical axis indicates the average reliability score. The reliability limits $\tau = 0.9$ and $\delta_r = 0.25$ are indicated with horizontal red dotted lines.

We can observe that $\tau \geq 0.9$ is reached with $N-1$ or $N-2$ subsampling for images, but not for videos. For videos, only the weighted variant τ_w barely reaches 0.9 at $N-1$ subsampling, indicating that video ranking was less reliable than images. In contrast, the relative sensitivity limit, which also takes into account the score values, is easily reached in all cases even at lower sampling sizes (at 50% sampling or even smaller). The only exception is 2017.Video data, where the limit is reached only at sampling 25 videos (83%). Finally, we can observe that both Kendall's scores tend to 1 and the relative sensitivity tends to 0 as the number of queries that are evaluated increases.

6 State-of-the-art deep neural networks

To account for current state-of-the-art deep neural network capabilities, we evaluate the performance of three recent image and video classification architectures, which

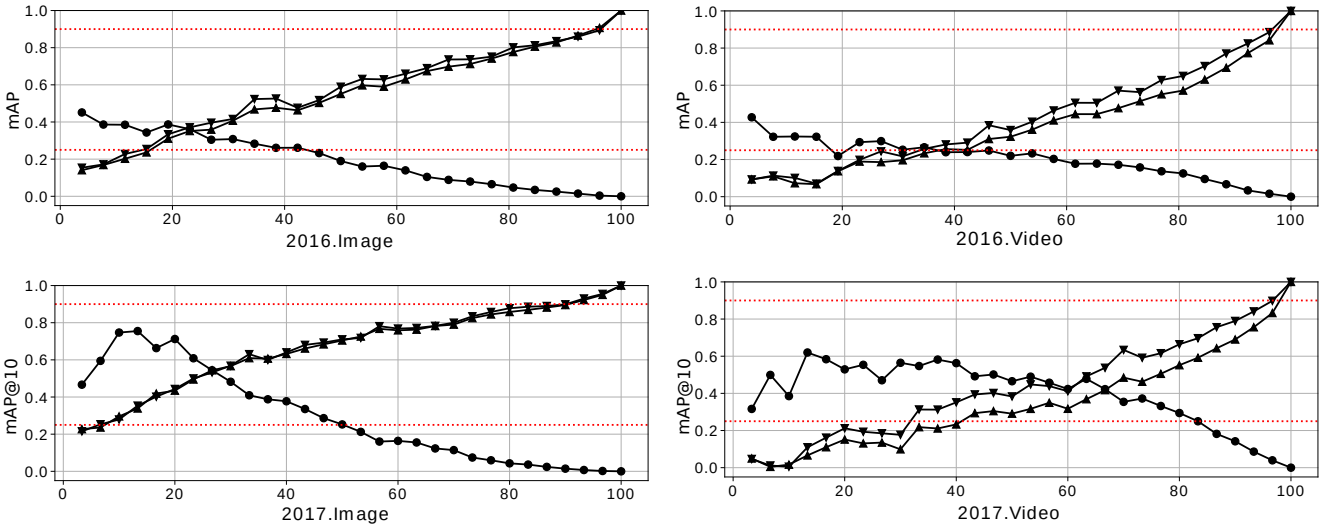

Fig. 9 Reliability scores of the system rankings: Year.Type represents the year of the data (2016 or 2017) and its type (Image or Video). X-axis is the subsampling percentage (sampling is performed at movie level) and y-axis is the reliability score. Relative sensitivity scores are marked with •, Kendall's tau with ▲, and weighted Kendall's with ▼. The reliability limits for the scores $\tau = 0.9$ and $\delta_\tau = 0.25$ are indicated with horizontal red dotted lines. For reference, at 100% subsampling, we trivially have perfect reliability as we would compare identical subsets.

were finetuned on the Interestingness10k data. We selected for the image data the ResNeXt-101-32x48d [91], PNASNet-5 [92], and ResNet-50 [93] architectures, augmented with best practices as presented in [94]; and for the video data, the GSM-InceptionV3 En3 [95], IR-CSN-152 [96], and R(2+1)-18 [97] architectures. The achieved results are synthesized Table 6.

Image classification. For image classification, we have followed the training protocol in [94]. In this context, we have fine-tuned all of the three algorithms using the provided weights trained on 940 million public images with 1.5k hashtags matching with 1,000 ImageNet1K synsets [98], fine-tuned on the ImageNet1K data set [99]. We adopt the set of good practices proposed by the authors, namely data augmentation including resizing the images, random horizontal shift of the center crop, horizontal flip and color jittering, including batch normalization layers, classification of the images at several resolutions and average the classification scores. The best results were achieved by FixResNeXt-101-32x48d, in both 2016, and 2017 scenarios, with a mAP and mAP@10 score of 0.2273 and 0.141, respectively.

Video classification. For video classification, we have fine-tuned the IR-CSN-152 and R(2+1)-18 networks, following the training protocol in [100] using the provided weights pre-trained on the IG-65M [100] data set, and fine-tuned on the Kinetics-400 [103] data set. We follow the good practices recommended by the authors which include a random patch cropping strategy, variable clip length, and temporal jittering. For GSM-InceptionV3 En3 [95], we followed the training protocol provided by the authors including the fusion of three variants of different clip lengths. The best results were achieved by GSM-InceptionV3 En3, with a mAP score of 0.1738 for the 2016 data, and a mAP@10 score of 0.0821, for the 2017 data.

Table 6 Performance of state-of-the-art deep neural network architectures when trained on the Interestingness10k data (bestME stands for best method from the MediaEval benchmark and bestSoA for the best method from the literature that was tested on these data).

	Method	2016 (mAP)	2017 (mAP@10)
Image	bestME	0.2336	0.1385
	bestSoA	0.2485	0.1560
	FixResNet50 [94]	0.1906	0.1099
	FixPNASNet-5 [94]	0.1981	0.1233
	FixResNeXt-101-32x48d [94]	0.2273	0.1410
Video	bestME	0.1815	0.0827
	bestSoA	0.1815	0.0930
	IR-CSN-152 [100]	0.1577	0.0629
	R(2+1)-18 [100]	0.1579	0.0644
	GSM-InceptionV3-En3 [95]	0.1738	0.0821

Overall analysis. The analysis of the results shows that these deep neural networks do not achieve the best results. While in a few cases, e.g., FixResNeXt-101, the best results from the MediaEval benchmark have been surpassed, none of the tested networks managed to surpass the current state of the art in media interestingness. Given the fact that the selected networks represent the current state-of-the-art in their corresponding domains, i.e., image and video classification tasks [16,

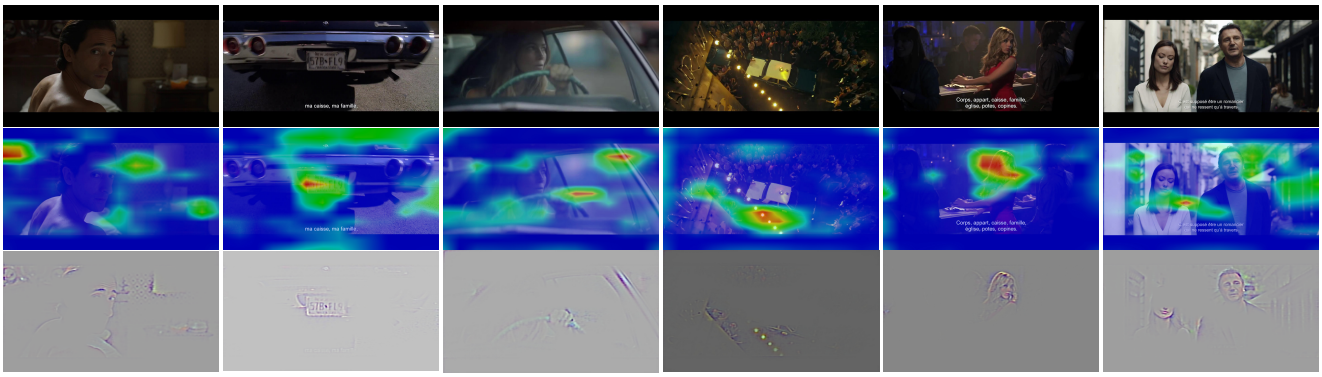

Fig. 10 Grad-CAM analysis of the network interpretation of visual information in the case of images predicted as interesting: the top row presents the original samples, the middle row presents the Grad-CAM output image describing "class-discriminative regions" [101], and the bottom row presents the Guided Backpropagation [102] Grad-CAM describing the features that most contributed to the class decision.

104], the intuition is that more specialised approaches are required to cope with this highly subjective task.

To understand how deep learning algorithms interpret the visual samples and thus how they attempt to predict interestingness, we computed the Grad-CAM maps via Grad-CAM [101] and Guided Backpropagation [102] Grad-CAM. Some relevant examples are presented in Figure 10. Results show that in many cases, the model focuses on the main subject, but predominantly more on elements adjacent to it, showing an inclination for detecting the context that surrounds the main subject. This is also true for human subjects, as the Grad-CAM analysis shows network activation on human faces, but also many times around the face. We theorize that this concentration of useful features on and around faces may represent a positive influence on the final results, as faces convey emotions.

7 Super-system design

In this final experiment, we investigate the possibility of exploiting the power of many systems to create a state-of-the-art performing super-system. The idea is to use an automatic, ad-hoc fusion strategy to exploit the advantages of each individual system. We prove that although individual systems are powerful, and declared state-of-the-art, there is always the possibility of achieving a greater performance via fusing system outputs. Though some of these state-of-the-art systems already include fusion strategies, our proposed ad-hoc fusion will incorporate the entire set of systems used during the MediaEval competition, therefore a larger set of system outputs. To achieve this goal, we investigate several standard approaches such as late fusion and boosting and, in the end, introduce a new fusion scheme based on a deep multilayer perceptron architecture with dense layers.

7.1 Evaluation setup

Ensembling requires typically tens of systems to be able to boost the performance. In practice, it is basically impossible to implement or retrieve such a number of systems from the authors, considering also re-running them in the very same conditions. There are also no best practices in this respect in the literature. The only approaches that do so use a very reduced number of inducers, e.g., less than 10 [105]. We therefore adopted a compromise that allows to use all the system runs submitted to the MediaEval benchmark, by experimenting solely on the *testset*. We use two split scenarios: (i) 75% training and 25% testing (*RSKF75*), and (ii) 50% training and 50% testing (*RSKF50*). Split samples are randomized and 100 partitions are generated. The official metrics are computed as average values over these partitions.

Although this approach looks more disadvantageous than training the systems on the entire *devset*, because the number of training items is significantly lower, we consider the results a good lower indicator of what the performance of late fusion would be. The following small experiment highlights the differences between the two training scenarios. We re-run our systems submitted to the MediaEval 2017 Interestingness task [78] under the new *testset* split conditions. As expected, results for the RSKF75 split are better than the ones for the RSKF50 split. However, the drop in performance is significant when compared with the original results attained by training on the entire *devset* and testing on *testset*. Thus, our system's mAP@10 results [78] decreased from 0.0555 (original devset/testset) to 0.0295 (RSKF75) for the image data, and from 0.0732 (original devset/testset) to 0.0314 (RSKF75) for the video data.

7.2 Approaches

We experiment with the following approaches: late fusion, boosting and proposed MLP-based architecture, which are presented in the next sections.

7.2.1 Late fusion

We investigate the possibility of using standard late fusion techniques, Kittler et al [106]. We experiment with producing an aggregated visual interestingness score via the minimum (*LFmin*), maximum (*LFmax*), mean (*LFmean*), and median (*LFmedian*) values of all interestingness scores of all the individual systems.

We also investigate a learning strategy via a weighted mean of system outputs (*LFweight*), where the weights are determined by the rank of the system in comparison with the other systems. Given that some systems may negatively affect the aggregated prediction, we use only the top–N systems, where N is empirically determined. The aggregated visual interestingness score is determined as $\sum_{i=1}^{K} w_i \cdot s_i$, where, for each individual sample, N is the total number of systems taken into account, w_i is the assigned weight for each system according to its rank, and s_i is the interestingness score. N is set to 2, 3, 5, 10, 20, and the number of systems. Weights are computed as $w_i = 1 - i * \alpha$, where α is varied between 0.01 and 0.5.

Overall, *LFweigh* had the best performance. For the *RSKF50* configuration, 2017.Video data represent the exception, where *LFmean* had better results, mAP@10 of 0.0872. *LFweigh* performed best in the following situations: on 2016.Image data, mAP of 0.2499 (using top $N = 10$ systems, $\alpha = 0.08$), on 2016.Video data, mAP of 0.1915 (using top $N = 10$ systems, $\alpha = 0.1$), and on 2017.Image data, mAP@10 of 0.1567 (using top $N = 20$ systems, $\alpha = 0.06$). For the *RSKF75* configuration, *LFmean* performed best on: 2016.Image data, mAP of 0.2519 (using top $N = 2$ systems, $\alpha = 0.25$), on 2016.Video, mAP of 0.1929 (using top $N = 10$ systems, $\alpha = 0.09$), on 2017.Image data, mAP@10 of 0.1532 (using top $N = 10$ systems, $\alpha = 0.11$), and finally on 2017.Video, mAP@10 of 0.0893 (using top $N = 10$ systems, $\alpha = 0.08$). While the use of late fusion combinations created systems that outperformed the MediaEval best results, in some cases, e.g., on 2017.Image and 2017.Video data, there are state-of-the-art systems that had better scores. Figure 11 presents the comparison of the best two performing late fusion systems with the other approaches.

7.2.2 Boosting

Boosting schemes are widely used for enhancing the performance of weak learners by aggregating them into a stronger classifier [107–109]. We experimented with several consecrated strategies, namely: AdaBoost, Freund and Schapire [110], and Gradient Boosting, Friedman [111]. We experimented with various combinations of systems based on their individual performance, from the worst performers to the best ones.

AdaBoost performed best under the *RSKF75* configuration on 2016.Image data, mAP of 0.2677 (aggregating systems ranked 8 to 10), on 2017.Image data, mAP@10 of 0.1674 (aggregating systems ranked 5 to 19), and on 2017.Video data, mAP@10 of 0.1129 (aggregating systems ranked 19 to 21). Under the *RSKF50* configuration, the best results are on 2016.Video data, mAP of 0.1987 (aggregating systems ranked 8 to 19). Gradient Boosting performed best under the *RSKF50* configuration on 2016.Image data, mAP of 0.2463 (aggregating systems ranked 1 to 20), on 2017.Video data, mAP@10 of 0.0961 (aggregating systems ranked 15 and 16). Under the *RSKF75* configuration, the best results are on 2016.Video data, mAP of 0.2209 (aggregating systems ranked 4 to 7). Overall, under the *RSKF75* configuration, boosting systems surpassed both the best MediaEval results and state-of-the-art results, while with the *RSKF50* configuration, there were better results from the state-of-the-art. Figure 11 presents the comparison of the best two performing boosting systems with the other approaches.

7.2.3 Proposed MLP architecture

We introduce a simple, yet efficient, fusion scheme that uses a deep MLP architecture. Our approach is motivated by the property of dense layers to at least weakly discover patterns and correlations between the individual systems decisions. We aim to model the bias learned by each system and the correlations between the biases to perform retrieval robustly and improve the overall performance of the aggregated system.

After experimenting with several architectures, we determined the following configuration: 10 layers, 5 dense layers (relu activation) with a batch normalization layer in-between each of them (totaling 4), inferring the final interestingness score with a single-layer linear perceptron (sigmoid activation). The architecture of the network is depicted in Figure 12.

In the training phase, the network takes as input the interestingness prediction scores of the systems to be aggregated, to learn complex joint decisions. All trainable weights of the networks are optimized together by

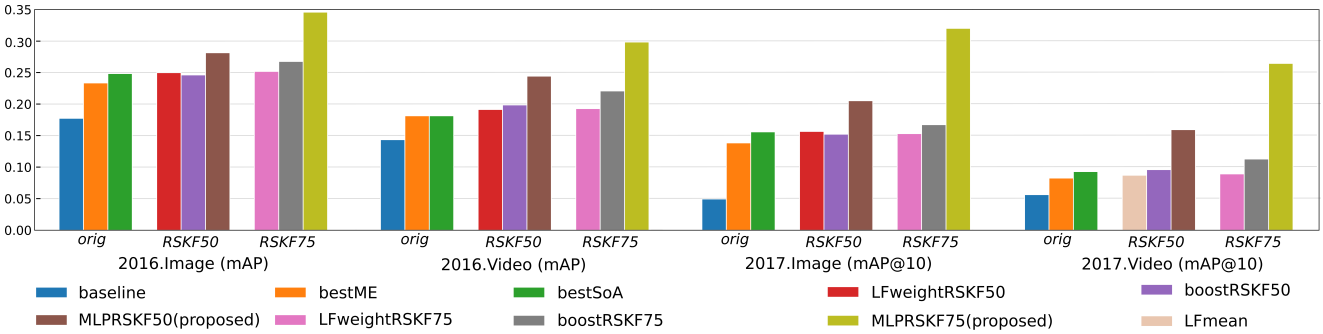

Fig. 11 Super-system design: *baseline* is a random ranking where samples are ranked randomly 5 times and mAP averaged, *bestME* and *bestSoA* are the best performers from the MediaEval benchmark and from the literature (in particular, are trained on the entire *devset*), respectively, *LF* stands for late fusion, *boost* for boosting, and *MLP* is the proposed Multi-Layer Perceptron scheme. We indicate the type of dataset split for the presented results: *orig* indicating the original split and *RSKF50* and *RSKF75* indicating the two generated splits. Results presented for RSKF50 and RSKF75 are computed as average values over 100 random partitions.

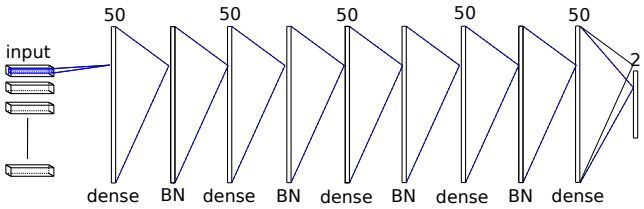

Fig. 12 Overview of the proposed MLP-based fusion scheme: 1 input layer followed by 4 pairs of dense/batch normalization (BN) layers, 1 dense layer, and 1 single-layer linear perceptron used for predicting the final interestingness score.

applying a stochastic gradient descent using the Adam approach in Kingma and Ba [112], with the following parameters: $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-08$. The loss function is set to the standard binary cross-entropy. The network was trained for 200 epochs with a batch size of 64. To simulate the benchmark scenario, we optimize the network according to accuracy and test it with the official benchmark metrics, while creating splits on the *testset* in both *RSKF50* and *RSKF75* configurations. Given a new set of images/videos, the network treats the input systems as untrained raters to model the common visual interestingness level shared between them.

We analyze the results obtained with the investigated and proposed MLP-based system and compare them with the best systems from the MediaEval benchmark and from literature. Results are summarized in Figure 11. Overall, clearly, the aggregated systems provide better results than the best individual systems. This is more or less expected given the fact that they exploit the advantages of several different systems. However, the proposed MLP-based learning strategy allows for a significant boost in performance. On 2016.Image and 2016.Video data, it improves the best results from a mAP of 0.2485 to 0.3459, and from 0.1815 to 0.2985, respectively. On 2017.Image and 2017.Video data, it improves the best results from a mAP@10 of 0.156 to 0.2646, and from 0.093 to 0.3202.

The improvement is dependent on the amount of training data used for the MLP, the top results being obtained for the *RSKF75* configuration. Nevertheless, good improvement is achieved in the *RSKF50* configuration as well.

Limitations. To understand the limitations of our approach, we empirically analyzed the results. We discuss here some of the common misclassification cases to understand the limitations of our approach. For certain types of visual samples, the inducers that we use as input into our fusion system display a correlated, positive or negative bias and the late fusion approach is not able to suppress this bias. Figure 13 illustrates some of the typical failure cases, i.e., false negative and false positive examples. For the false negative examples, we observe a number of darker interesting images that are incorrectly classified as non-interesting, with their interestingness score often being lower than 0.1. This may be the result of inducer algorithms not having enough visual information to correctly score these particular samples. For the false positive examples, some outdoor non-interesting images (compared to their class representatives), usually containing groups of people, are assigned a high interestingness score, typically greater than 0.5. This may represent an indication that the inducer algorithms tend to pay more attention to visual samples that contain people and therefore present a bias for those particular cases.

8 Conclusions and open questions

The prediction of *visual interestingness* is a research topic of increasing importance in the multimedia com-

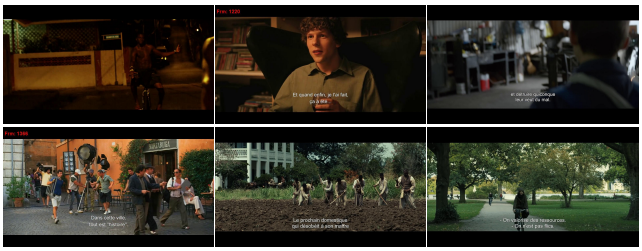

Fig. 13 Failure case examples: top row presents false negatives (FN), and bottom row presents false positives (FP).

munity, with practical applications in advertising, social media, education, media recommendation and many more. In this work, we introduced a publicly available, common evaluation framework for image and video visual interestingness prediction. It consists of a robust data set, with 9,831 images and more than 4 hours of video, and interestingness scores determined from over 1M pair-wise annotations of 800 trusted annotators.

To account for baseline systems, we provide an in-depth analysis of the crucial components of visual interestingness prediction by reviewing the capabilities and the evolution of 192 validated systems (129 from the MediaEval benchmark and 63 state-of-the-art systems from the literature). We analyze overall capabilities, influence of the employed features and techniques, and generalization capabilities. For the 129 ranked systems of the MediaEval benchmark, computed relative sensitivity, Kendall's rank correlation, and weighted Kendall's rank show good reliability of the results. We also discuss the possibility of going beyond state-of-the-art performance via an automatic, ad-hoc system fusion, and propose a deep MLP-based architecture that outperforms the state-of-the-art systems by a large margin.

We summarize below the most important lessons learned and insights gained, as well as identify the remaining open questions and perspectives.

8.1 What are the current capabilities?

Overall system performance. Over the analyzed systems, without taking into account our system fusion experiments, the highest precision for image visual interestingness prediction is obtained via a learning-to-rank DNN using both deep features and a deep ranking approach, Parekh et al. [52], mAP of 0.3125. For video, the highest precision is achieved via a late fusion between a learning-to-rank Siamese network and a reinforcement ranking based on a Markov decision process, using both visual and audio descriptors, Wang et al [54], mAP of 0.2228. Globally speaking, results are not that high compared to other classification and regression tasks, and are similar to the ones achieved in early object classification, e.g., see early TRECVid campaigns. Naturally, video prediction is more challenging than image prediction, as results show. Data, annotations and techniques should still adapt and come with new improvements to address this subjective task. Nevertheless, one should notice an ascending trend, as performance significantly increased over the years, e.g., in 2016 best mAP is 0.2336 for images, Liem [64], and 0.1815 for videos, Alemida [53]; in 2017 it reaches 0.3075 for images, Permadi et al. [73], and 0.2094 for videos, Ben-Ahmed et al. [65]; and in 2018 0.3125 for images, Parekh et al. [52], and 0.2228 for videos, Wang et al. [54]. Therefore, progress is continuously made.

Content representation and methods. It is interesting to see the rich diversity of approaches, from data representation to the prediction methods. The most popular content description categories are *unimodal* representations, accounting for 72% of the analyzed methods, followed by *deep features* (employed alone or as part of multimodal and fusion systems) with 59%. The most popular methods are *SVMs* used as classifiers, accounting for 30% of the analyzed methods, followed by *DNNs* with 28%, and *ranking* techniques with 13%. The best performing system is, of course, a combination of the two, i.e., description scheme and prediction approach. Some of the image prediction best systems are proposed by Parekh et al. [52]. They use a unimodal approach via AlexNet fc7 layer features and a learning-to-rank DNN, mAP of 0.3125. For video, the best systems are the approaches of Wang et al. [54]. They use, either LBP-based features alone, or early fusion of deep features extracted from InceptionV3, AlexNet and C3D, and traditional visual and audio features, with Siamese networks, achieving a mAP from 0.2131 to 0.2228. We would like to also highlight the best performing SVMs for images, i.e., Permadi et al. [73], via a polynomial kernel SVM, mAP of 0.3052, and for video, Ben-Ahmed et al. [65], via a linear kernel SVM, mAP of 0.2122. The best performing ranking approaches for images are: Almeida and Savii [79], via RankBoost, mAP of 0.271, and for video, Almeida and Savii [79], via rankSVM, mAP of 0.1877. It is worth noting that, depending on the data, state-of-the-art results are not necessarily obtained using deep learning, although is predominant.

Generalization capabilities. Annotated data is scarce to fill in the requirements of current deep neural networks. Regardless of the efforts of releasing more and more annotated data, it is not a sustainable action in the mid term. Systems have to find alternate solutions for training the algorithms. Unsupervised techniques, although very appealing, are still too incipient for this type of subjective tasks. A viable immediate alternative is to borrow data from adjacent domains

and use transfer learning techniques. We noticed an encouraging trend in this direction. 45% of the analyzed systems used at least a *pre-trained extraction* generalization scheme, i.e., systems are trained on data unrelated to interestingness, like object recognition data sets, and used directly, usually as features in a classifier, to predict interestingness. These systems were the overall state-of-the-art performers. 9% of the systems went further, and use *fine-tuning* approaches, i.e., systems are firstly trained on data unrelated to interestingness and then retrained on the Interestingness10k data. For image prediction, Erdogan et al. [69] obtains the best performance via a fine-tuned AlexNet network, with a mAP of 0.2125. For video prediction, Ben-Ahmed et al. [65] obtains the best performance via a video genre classification system, with a mAP of 0.2094. Significantly less, 3% represent *correlated* approaches, via systems trained on external data, from correlated domains, e.g., memorability, aesthetics, which are used to predict interestingness. For image prediction the best approach uses a social interestingness prediction system trained on Flickr data, Shen et al. [71], mAP of 0.2336. For video prediction, Xu et al. [61] use SentiBank features, based on emotional content, achieving a mAP of 0.154.

Ad-hoc fusion. Another important observation is the fact that regardless how good a system is, the fusion of the results from several systems, even with individual average performance, proves to increase the performance. After experimenting with several fusion techniques, like standard late fusion of system scores, boosting techniques that use weak learners and a proposed, deep MLP-based system fusion, we were able to boost performance almost in every situation. The proposed MLP system achieves a maximum improvement of 105% on image prediction over state-of-the-art results, improving mAP from 0.156 to 0.3202, and of 184% on video prediction, improving mAP from 0.093 to 0.2646. The inherent disadvantage is the significantly higher computational complexity of the aggregated system. However, good performance was obtained by fusing few systems, an order of ten, e.g., 30-40. With current hardware acceleration and parallel computing, this is a feasible alternative.

Recommendations to system performance. During our analysis, some approaches stood out when compared with the others. For instance, when analyzing modalities, deep and traditional visual features show promising results. However, a more obvious outlier is represented by late fusion systems. On average, the performance of such systems was better, both for image and for video data (average mAP over the analyzed systems of 0.2416 and 0.1878, respectively). This observation is enforced by the good performance of hybrid classifiers on video data, that use more than one type of classifier (as presented in Section 5.3.2) but also by the top performance of our proposed late fusion MLP system. The intuition is that this may be an effect of the inherent subjective and multi-modal aspects of interestingness. Furthermore, while deep learning-based systems do not necessarily represent the state-of-the-art performers, they do present some interesting results. For instance, when analyzing the average performance of the method categories, deep neural networks achieved the highest average score for image data. Regarding the performance of modern DNN approaches, tested in Section 6, while these methods do not outperform the state of the art, some of these networks, such as GSM-InceptionV3-En3 [95] and FixResNeXt-101-32x48d [94], achieve very high scores. Finally, some good training practices are studied in Sections 5.3.2 and 5.4.1. For instance, when extracting features from a deep semantic embedding model, Vasudevan et al. [82], achieves better results when finetuning the semantic model with Interestingness10k data, as opposed to directly extracting the embeddings. Other good practices involve using external data from correlated domains like social interestingness and emotional content. This type of data augmentation, paired with data upsampling on Interestingness10k images contributed, for example, to the best mAP score on 2016.Image data achieved during the MediaEval competition [71].

8.2 What are the open questions remaining?

System performance. Although a great deal of methods were experimented with various feature representations, fusion techniques and transfer learning, top performance is just around a mAP of 0.31% and 0.22%, for image and video prediction, respectively. Current performance on video prediction is significantly lower than for images. This is still incipient and requires significant improvements. At annotations level, a lead is to deepen the understanding of the concept of interestingness and visual information by exploring more related subjective concepts. Psychological user studies revealed many concepts related to interestingness that have great potential in improving its understanding, e.g., *novelty*, *coping potential*, *complexity*, *comprehensibility*. Interestingness prediction is a multifaceted problem and should be approached from a more interconnected perspective. For a comprehensive analysis of the correlation between interestingness and other concepts, from the psychological, experimental and computer vision points of view, we refer the reader to Constantin et al. [4]. At the methods level, *temporal information* remains largely unexplored

for video prediction. Therefore, a future lead is to augment prediction models using *temporal-based models*, whether they are based on new DNN architectures or on *temporal aggregation of features*, for better encoding of video information. Another lead is to explore the *attention mechanism* in DNN architectures, so as to focus the interestingness prediction on certain regions of the image and video. A small region in the image may raise great interest to the viewer, rather than the whole image itself.

Ground truth data. Another open challenge is the generation of meaningful training data. Deep learning models proved again to be state-of-the-art performers, therefore, there is the need of more annotated data. Given the subjectivity of the task, the annotation is not as straightforward as for example, for object annotation. Everybody understands what a chair or a tree looks like, but what is interesting is not the same for everybody. This is clearly visible in the Interestingness10k annotations. Although we used expert annotators, i.e., human assessors that were given thorough guidance on the task and scientific problem, the annotator agreement was average to good, with a kappa value of 0.556 and 0.519, for images and videos, respectively. The annotation mechanism, e.g., pair-wise comparisons, user studies, especially for videos, should be more investigated and, again, perhaps explored in correlation with other subjective properties.

Unsupervised learning. Unsupervised generation of data has currently proved a feasible task for many classification systems. Significant progress has been made via auto-encoders and generative adversarial networks (GAN). However, it was still not explored for the generation of images according to their perception. The closest experiments are for generating human faces with different emotions. This would be a pioneering direction to explore, i.e., training GANs to automatically generate data with different levels of interestingness.

Acknowledgements We would like to acknowledge first, Technicolor France for founding and supporting the Interestingness10k data set and the Predicting Media Interestingness task. We acknowledge the work of our fellow task coorganizers (in alphabetical order): Alexey Ozerov, Frédéric Lefebvre, Hanli Wang, Michael Gygli, Toan Do, Vincent Demoulin, and Yu-Gang Jiang. We would like to acknowledge also the MediaEval Benchmarking Initiative for Multimedia Evaluation and in particular Martha Larson, for hosting the Predicting Media Interestingness Task, constant support and enlightening discussions.

The work of Mihai Gabriel Constantin and Bogdan Ionescu was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, via agreement 2SOL/2017, and from project AI4Media, A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911. The work of Liviu-Daniel Ştefan was supported by the Operational Programme Human Capital of the Ministry of Europe Funds through the Financial Agreement 51675/ 09.07.2019, SMIS code 125125.

References

1. H. Squalli-Houssaini, N. Q. K. Duong, M. Gwenaëlle, and C.-H. Demarty, "Deep learning for predicting image memorability," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2371–2375, IEEE, 2018.
2. S. Mo, J. Niu, Y. Su, and S. K. Das, "A novel feature set for video emotion recognition," *Neurocomputing*, vol. 291, pp. 11–20, 2018.
3. A. Carballal, C. Fernandez-Lozano, J. Heras, and J. Romero, "Transfer learning features for predicting aesthetics through a novel hybrid machine learning method," *Neural Computing and Applications*, pp. 1–12, 2019.
4. M. G. Constantin, M. Redi, G. Zen, and B. Ionescu, "Computational understanding of visual interestingness beyond semantics: Literature survey and analysis of covariates," *ACM Computing Surveys*, 2019.
5. D. E. Berlyne, "Interest as a psychological concept," *British journal of psychology. General section*, vol. 39, no. 4, pp. 184–195, 1949.
6. D. E. Berlyne, *Conflict, arousal, and curiosity.* McGraw-Hill Book Company, 1960.
7. D. E. Berlyne, "Novelty, complexity, and hedonic value," *Perception & Psychophysics*, vol. 8, no. 5, pp. 279–286, 1970.
8. S. Hidi and V. Anderson, "Situational interest and its impact on reading and expository writing," *The role of interest in learning and development*, vol. 11, pp. 213–214, 1992.
9. C. Chamaret, C.-H. Demarty, V. Demoulin, and G. Marquant, "Experiencing the interestingness concept within and between pictures," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–12, 2016.
10. P. J. Silvia, "What is interesting? exploring the appraisal structure of interest.," *Emotion*, vol. 5, no. 1, p. 89, 2005.
11. P. J. Silvia, "Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions.," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 3, no. 1, p. 48, 2009.
12. R. R. McCrae, "Aesthetic chills as a universal marker of openness to experience," *Motivation and Emotion*, vol. 31, no. 1, pp. 5–11, 2007.
13. L.-C. Hsieh, W. H. Hsu, and H.-C. Wang, "Investigating and predicting social and visual image interestingness on social media by crowdsourcing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4309–4313, IEEE, 2014.
14. C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre, "Mediaeval 2016 predicting media interestingness task," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
15. M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

16. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
17. A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of trecvid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.
18. G. Awad, P. Over, and W. Kraaij, "Content-based video copy detection benchmarking at trecvid," *ACM Transactions on Information Systems (TOIS)*, vol. 32, no. 3, p. 14, 2014.
19. T. Deselaers, T. M. Deserno, and H. Müller, "Automatic medical image annotation in imageclef 2007: Overview, results, and discussion," *Pattern Recognition Letters*, vol. 29, no. 15, pp. 1988–1995, 2008.
20. J. Poignant, H. Bredin, and C. Barras, "Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22547–22567, 2017.
21. J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, and H. Müller, "Evaluating performance of biomedical image retrieval systemsan overview of the medical image retrieval task at imageclef 2004–2013," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 55–61, 2015.
22. M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1633–1640, IEEE, 2013.
23. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.
24. M. Soleymani, "The quest for visual interest," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 919–922, ACM, 2015.
25. Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1–7, 2013.
26. H. Grabner, F. Nater, M. Druey, and L. Van Gool, "Visual interestingness in image sequences," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 1017–1026, ACM, 2013.
27. M. Gygli and M. Soleymani, "Analyzing and predicting gif interestingness," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 122–126, ACM, 2016.
28. M. Gygli, Y. Song, and L. Cao, "Video2gif: Automatic generation of animated gifs from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1001–1009, IEEE, 2016.
29. S. Bakhshi, D. A. Shamma, L. Kennedy, Y. Song, P. De Juan, and J. Kaye, "Fast, cheap, and good: Why animated gifs engage us," in *Proceedings of the chi conference on human factors in computing systems*, pp. 575–586, ACM, 2016.
30. C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, M. Gygli, and N. Q. K. Duong, "Mediaeval 2017 predicting media interestingness task," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
31. C.-H. Demarty, M. Sjöberg, M. G. Constantin, N. Q. Duong, B. Ionescu, T.-T. Do, and H. Wang, "Predicting interestingness of visual content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*, pp. 233–265, Cham: Springer, 2017.
32. P. Salesses, K. Schechtner, and C. A. Hidalgo, "The collaborative image of the city: mapping the inequality of urban perception," *PloS one*, vol. 8, no. 7, 2013.
33. Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.
34. S. Ovadia, "Ratings and rankings: reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.
35. G. N. Yannakakis and J. Hallam, "Ranking vs. preference: a comparative study of self-reporting," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 437–446, Springer, 2011.
36. J. Li, M. Barkowsky, and P. L. Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3dtv: performances of three different designs," in *Proceedings of SPIE Electronic Imaging, Stereoscopic Displays and Applications*, vol. 8648, 2013.
37. R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: the method of paired comparisons," *Biometrika*, no. 39 (3-4), pp. 324–345, 1952.
38. A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1.1, pp. 77–89, 2007.
39. J. J. Randolph, "Free-marginal multirater kappa (multirater free): an alternative to fleiss' fixed-marginal multirater kappa," in *Joensuu Learning and Instruction Symposium, Joensuu, Finland*, 2005.
40. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
41. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition*, vol. 1, pp. 886–893, IEEE Computer Society, 2005.
42. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 971–987, 2002.
43. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
44. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, IEEE, 2015.
45. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, IEEE, 2006.
46. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, IEEE, 2010.
47. Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super fast event recognition in internet videos," *IEEE*

Transactions on Multimedia, vol. 17, no. 8, pp. 1174–1186, 2015.
48. M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5*, BMVA Press, 2014.
49. C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," *SIGIR Forum*, vol. 51, no. 2, pp. 235–242, 2017.
50. H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.
51. M. G. Constantin and B. Ionescu, "Content description for predicting image interestingness," in *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pp. 1–4, IEEE, 13–14 July 2017.
52. J. Parekh, H. Tibrewal, and S. Parekh, "Deep pairwise classification and ranking for predicting media interestingness," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR, Yokohama, Japan, June 11-14.*, pp. 428–433, ACM, 2018.
53. J. Almeida, "UNIFESP at mediaeval 2016: Predicting media interestingness task," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
54. S. Wang, S. Chen, J. Zhao, and Q. 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*, ASMMC-MMAC'18, pp. 55–61, ACM, 2018.
55. R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European conference on computer vision*, pp. 288–301, Springer, 2006.
56. C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE Journal of selected topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.
57. Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 419–426, IEEE, 2006.
58. J. Almeida, L. P. Valem, and D. C. Pedronette, "A rank aggregation framework for video interestingness prediction," in *International Conference on Image Analysis and Processing*, pp. 3–14, Springer, 2017.
59. J. Almeida, N. J. Leite, and R. d. S. Torres, "Comparison of video sequences with histograms of motion patterns," in *18th IEEE International Conference on Image Processing*, pp. 3673–3676, IEEE, 2011.
60. D. Borth, T. Chen, R. Ji, and S.-F. Chang, "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 459–460, ACM, 2013.
61. B. Xu, Y. Fu, and Y. Jiang, "Bigvid at mediaeval 2016: Predicting interestingness in images and videos," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21*, vol. 1739, CEUR-WS.org, 2016.
62. R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.
63. E. Berson, C. Demarty, and N. Q. K. Duong, "Multimodality and deep learning when predicting media interestingness," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
64. C. Liem, "TUD-MMC at mediaeval 2016: Predicting media interestingness task," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
65. O. B. Ahmed, J. Wacker, A. Gaballo, and B. Huet, "Eurecom@mediaeval 2017: Media genre inference for predicting media interestingness," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
66. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
67. Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, December 5-10, Barcelona, Spain*, pp. 892–900, 2016.
68. K. Sivaraman and G. Somappa, "Moviescope: Movie trailer classification using deep neural networks," *University of Virginia*, 2016.
69. G. Erdogan, A. Erdem, and E. Erdem, "HUCVL at mediaeval 2016: Predicting interesting key frames with deep models," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
70. V. Lam, T. Do, S. Phan, D.-D. Le, S. Satoh, and D. A. Duong, "Nii-uit at mediaeval 2016 predicting media interestingness task.," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
71. Y. Shen, C. Demarty, and N. Q. K. Duong, "Technicolor@mediaeval 2016 predicting media interestingness task," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
72. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678, ACM, 2014.
73. R. A. Permadi, S. G. P. Putra, Helmiriawan, and C. C. S. Liem, "DUT-MMSR at mediaeval 2017: Predicting media interestingness task," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
74. F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM, 2010.
75. S. Rayatdoost and M. Soleymani, "Ranking images and videos on visual interestingness by visual sentiment features.," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
76. Y. Liu, Z. Gu, and T. H. Ko, "Predicting media interestingness via biased discriminant embedding and supervised manifold regression.," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
77. Y. Liu, Z. Gu, T. H. Ko, and K. A. Hua, "Learning perceptual embeddings with two related tasks for joint predictions of media interestingness and emotions," in *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pp. 420–427, ACM, 2018.

78. M. G. Constantin, B. A. Boteanu, and B. Ionescu, "Lapi at mediaeval 2017-predicting media interestingness.," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
79. J. Almeida and R. M. Savii, "GIBIS at mediaeval 2017: Predicting media interestingness task," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, CEUR-WS.org, 2017.
80. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
81. Y. Shen, C.-H. Demarty, and N. Q. K. Duong, "Deep learning for multimodal-based video interestingness prediction," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1003–1008, IEEE, 2017.
82. A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool, "Eth-cvl@ mediaeval 2016: Textual-visual embeddings and video2gif for video interestingness," in *MediaEval Workshop, Hilversum, The Netherlands, October 20-21.*, vol. 1739, CEUR-WS.org, 2016.
83. X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li, "Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 243–252, 2013.
84. A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2390–2398, 2015.
85. F. Liu, Y. Niu, and M. Gleicher, "Using web photos for measuring video frame interestingness," in *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.
86. J. Urbano, M. Marrero, and D. Martín, "On the measurement of test collection reliability," in *The 36th International ACM SIGIR conference on research and development in Information Retrieval*, pp. 393–402, ACM, July 28 - August 1 2013.
87. M. Sanderson and J. Zobel, "Information retrieval system evaluation: effort, sensitivity, and reliability," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169, ACM, August 15-19 2005.
88. H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics. Sage*, pp. 508–510, 2007.
89. E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, eds.), pp. 315–323, ACM, August 24-28 1998.
90. S. Vigna, "A weighted correlation index for rankings with ties," in *Proceedings of the 24th International Conference on World Wide Web, WWW* (A. Gangemi, S. Leonardi, and A. Panconesi, eds.), pp. 1166–1176, ACM, May 18-22 2015.
91. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.
92. C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 19–34, 2018.
93. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
94. H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," in *Advances in Neural Information Processing Systems*, pp. 8250–8260, 2019.
95. S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
96. D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5552–5561, 2019.
97. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.
98. I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.
99. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
100. D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12046–12055, 2019.
101. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.
102. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
103. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
104. R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The" something something" video database for learning and evaluating visual common sense.," in *ICCV*, vol. 1, p. 5, 2017.
105. X. Li, Y. Huo, Q. Jin, and J. Xu, "Detecting violence in video using subclasses," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016*, pp. 586–590, ACM, October 15-19 2016.
106. J. Kittler, M. Hater, and R. P. Duin, "Combining classifiers," in *Proceedings of 13th international conference on pattern recognition*, vol. 2, pp. 897–901, IEEE, 1996.
107. S. Han, Z. Meng, A.-S. Khan, and Y. Tong, "Incremental boosting convolutional neural network for facial action unit recognition," in *Advances in neural information processing systems*, pp. 109–117, 2016.
108. M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Bier-boosting independent embeddings robustly," in

Proceedings of the IEEE International Conference on Computer Vision, pp. 5189–5198, 2017.
109. J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3056–3064, 2015.
110. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
111. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
112. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, Conference Track Proceedings*, 2015.