Predicting Interestingness of Visual Content

Claire-Helène Demarty, Mats Sjöberg, Gabriel Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang

Predicting Interestingness of Visual Content

Claire-Hélène Demarty, Mats Sjöberg, Mihai Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do and Hanli Wang

Abstract The ability of multimedia data to attract and keep people's interest for longer periods of time is gaining more and more importance in the fields of information retrieval and recommendation, especially in the context of the ever growing market value of social media and advertising. In this chapter we introduce a benchmarking framework (dataset and evaluation tools) designed specifically for assessing the performance of media interestingness prediction techniques. We release a dataset which consists of excerpts from 78 movie trailers of Hollywood-like movies. These data are annotated by human assessors according to their degree of interestingness. A real-world use scenario is targeted, namely interestingness is defined in the context of selecting visual content for illustrating a Video on Demand (VOD) website. We provide an in-depth analysis of the human aspects of this task, *i.e.,* the correlation between perceptual characteristics of the content and the actual data, as well as of the machine aspects by overviewing the participating systems of

Claire-Hélène Demarty
Technicolor R&I, France, e-mail: claire-helene.demarty@technicolor.com

Mats Sjöberg
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland, e-mail: mats.sjoberg@helsinki.fi

Mihai Gabriel Constantin
LAPI, University Politehnica of Bucharest, Romania, e-mail: mgconstantin@imag.pub.ro

Ngoc Q. K. Duong
Technicolor R&I, France, e-mail: quang-khanh-ngoc.duong@technicolor.com

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, Romania, e-mail: bionescu@imag.pub.ro

Thanh-Toan Do
Singapore University of Technology and Design, Singapore & University of Science, Vietnam, e-mail: thanhtoan_do@sutd.edu.sg

Hanli Wang
Department of Computer Science and Technology, Tongji University, China, e-mail: hanliwang@tongji.edu.cn

the 2016 MediaEval Predicting Media Interestingness campaign. After discussing the state-of-art achievements, valuable insights, existing current capabilities as well as future challenges are presented.

1 Introduction

With the increased popularity of amateur and professional digital multimedia content, accessing relevant information is now dependent on effective tools for managing and browsing, due to the huge amount of data. Managing content often involves filtering parts of it to extract what corresponds to specific requests or applications. Fine filtering is impossible however without a clear understanding of the content's semantic meaning. To this end, current research in multimedia and computer vision has moved towards modeling of more complex semantic notions, such as emotions, complexity, memorability and interestingness of content, thus going closer to human perception.

Being able to assess, for instance, the interestingness level of an image or a video has several direct applications: from personal and professional content retrieval, content management, to content summarization and story telling, selective encoding, or even education. Although it has already raised a huge interest in the research community, a common and clear definition of multimedia interestingness has not yet been proposed, nor does a common benchmark for the assessment of the different techniques for its automatic prediction exist.

MediaEval[1] is a benchmarking initiative which focuses on the multi-modal aspects of multimedia content, *i.e.,* it is dedicated to the evaluation of new algorithms for multimedia access and retrieval. MediaEval emphasizes the multi-modal character of the data, *e.g.,* speech, audio, visual content, tags, users and context. In 2016, the Predicting Media Interesting Task[2] was proposed as a new track in the MediaEval benchmark. The purpose of the task is to answer a real and professional-oriented interestingness prediction use case, formulated by Technicolor[3]. Technicolor is a creative technology company and a provider of services in multimedia entertainment and solutions, in particular, providing also solutions for helping users select the most appropriate content according to, for example, their profile. In this context, the selected use case for interestingness consists in helping professionals to illustrate a Video on Demand (VOD) web site by selecting some interesting frames and/or video excerpts for the posted movies.

Although the targeted application is well-defined and confined to the illustration of a VOD web site, the task remains highly challenging. Firstly, it raises the question of the subjectivity of interestingness, which may vary from one person to the other. Furthermore, the semantic nature of interestingness constrains its modeling to

[1] http://www.multimediaeval.org/

[2] http://www.multimediaeval.org/mediaeval2016/mediainterestingness/

[3] http://www.technicolor.com

be able to bridge the semantic gap between the notion of interestingness and the statistical features that can be extracted from the content. Lastly, by placing the task in the field of the understanding of multi-modal content, *i.e.,* audio and video, we push the challenge even further by adding a new dimensionality to the task. The choice of Hollywood movies as targeted content also adds potential difficulties, in the sense that the systems will have to cope with different movie genres and potential editing and special effects (*i.e,* alteration of the content).

Nevertheless, although highly challenging, the building of the task responds to the absence of such benchmarks. It provides a common dataset and a common definition of interestingness. To the best of our knowledge, the MediaEval 2016 Predicting Media Interestingness is the first attempt to cope with this issue in the research community. Even though still in its infancy, the task has, in this first year, been a source of meaningful insights for the future of the field.

This chapter focuses on a detailed description of the benchmarking framework, together with a thorough analysis of its results, both in terms of the performance of the submitted systems and in what concerns the produced annotated dataset. We identify the following main contributions:

- an overview of the current interestingness literature, both from the perspective of the psychological implications and also from the multimedia/computer vision side;
- the introduction of the first benchmark framework for the validation of the techniques for predicting the interestingness of video (image and audio) content, formulated around a real-world use case, which allows for disambiguating the definition of interestingness;
- the public release of a specially designed annotated dataset. It is accompanied with an analysis of its perceptual characteristics;
- an overview of the current capabilities via the analysis of the submitted runs;
- an in-depth discussion on the remaining issues and challenges for the prediction of the interestingness of content.

The rest of the chapter is organized as follows. Section 2 presents a consistent state of the art on interestingness prediction from both the psychological and computational points of view. It is followed by a detailed description of the MediaEval Predicting Media Interestingness Task, its definition, dataset, annotations and evaluation rules, in Section 3. Section 4 gives an overview of the different submitted systems and trends for this first year of the benchmark. We analyze the produced dataset and annotations, their qualities and limitations. Finally, Section 5 discusses the future challenges and the conclusions.

2 A review of the literature

The prediction and detection of multimedia data interestingness has been analyzed in the literature from the human perspective, involving psychological studies, and

also from the computational perspective, where machines are taught to replicate the human process. Content interestingness has gained importance with the increasing popularity of social media, on-demand video services and recommender systems. These different research directions try to create a general model for human interest, go beyond the subjectivity of interestingness and detect some objective features that appeal to the majority of subjects. In the following, we present an overview of these directions.

2.1 Visual interestingness as a psychological concept

Psychologists and neuroscientists have extensively studied the subjective perception of visual content. The basis of the psychological interestingness studies was established in [5]. It was revealed that interest is determined by certain factors and their combinations, like "*novelty*", "*uncertainty*", "*conflict*" and "*complexity*". More recent studies have also developed the idea that interest is a result of appraisal structures [59]. Psychological experiments determined two components, namely: "*novelty-complexity*" — a structure that indicates the interest shown for new and complex events; and "*coping potential*" — a structure that measures a subject's ability to discern the meaning of a certain event. The influence of each appraisal component was further studied in [60], proving that personality traits could influence the appraisals that define interest. Subjects with a high "*openness*" trait, who are sensation seeking, curious, open to experiences [48], were more attracted by the novelty-complexity structure. In opposition, those not belonging to that personality category, were influenced more by their coping potential. Some of these factors were confirmed in numerous other studies based on image or video interestingness [11, 55, 22, 62].

The importance of objects was also analyzed as a central interestingness cue [20, 63]. The saliency maps used by the authors in [20] were able to predict interesting objects in a scene with an accuracy of more than 43%. They introduced and demonstrated the idea that, when asked to describe a scene, humans tend to talk about the most interesting objects in that scene first. Experiments show that there was a strong consistency between different users [63]. Eye movement, another behavioral cue, was used by the authors in [9] to detect the level of interest shown in segments of images or whole images. The authors used saccades, the eye movements that continuously contribute to the building of a mental map of the viewed scene. The authors in [4] studied the object attributes that could influence importance and draw attention, and found that animated, unusual or rare events tend to be more interesting for the viewer.

In [34], the authors conducted an interestingness study on 77 subjects, using artworks as visual data. The participants were asked to give ratings on different scales to opposing attributes for the images, including: "*interesting-uninteresting*", "*enjoyable-unenjoyable*", "*cheerful-sad*", "*pleasing-displeasing*". The results show that disturbing images can still be classified as interesting, therefore negating the

need of pleasantness in human visual interest stimulation. Another analysis [11] led to several conclusions regarding the influence on interest, namely: instant enjoyment was found to be an important factor, exploration intent and novelty had a positive effect and challenge had a small effect. The authors in [13] studied the influence of familiarity with the presented image on the concept of interestingness. They concluded that for general scenes, unfamiliar context positively influenced interest, while photos of familiar faces (including self photos) were more interesting than those of unfamiliar people.

It is interesting to observe also a correlation between different attributes and interestingness. Authors in [23] performed such a study on a specially designed and annotated dataset of images. The positively correlated attributes were found to be "*assumed memorability*", "*aesthetics*", "*pleasant*", "*exciting*", "*famous*", "*unusual*", "*makes happy*", "*expert photo*", "*mysterious*", "*outdoor-natural*", "*arousing*", "*strange*", "*historical*" or "*cultural place*".

2.2 Visual interestingness from a computational perspective

Besides the vast literature of psychological studies, the concept of visual interestingness has been studied from the perspective of automatic, machine-based, approaches. The idea is to replicate human capabilities via computational means.

For instance, the authors in [23] studied a large set of attributes: RGB values, GIST features [51], spatial pyramids of SIFT histograms [40], colorfulness [17], complexity, contrast and edge distributions [36], arousal [47] and composition of parts [6] to model different cues related to interestingness. They investigated the role of these cues in varying context of viewing: different datasets were used, from arbitrary selected and very different images (weak context) to images issued from similar Webcam streams (strong context). They found that the concept of "*unusualness*", defined as the degree of novelty of a certain image when compared to the whole dataset, was related to interestingness, in case of a strong context. Unusualness was calculated by clustering performed on the images using Local Outlier Factor [8] with RGB values, GIST and SIFT as features, composition of parts and complexity interpreted as the JPEG image size. In case of a weak context, personal preferences of the user, modeled by pixel values, GIST, SIFT and Color Histogram as features, classified with a ν-SVR — Support Vector Regression (SVR) with a RBF kernel, performed best. Continuing this work, the author in [62] noticed that a regression with sparse approximation of data performed better with the features defined by [23] than the SVR approach.

Another approach [19] selected three types of attributes for determining image interestingness: compositional, image content and sky-illumination. The compositional attributes were: rule of thirds, low depth of field, opposing colors and salient objects; the image content attributes were: the presence of people, animals and faces, indoor/outdoor classifiers; and finally the sky-illumination attributes consisted of scene classification as cloudy, clear or sunset/sunrise. Classification of interesting

content is performed with Support Vector Machines (SVM). As baseline, the authors used the low-level attributes proposed in [36], namely average hue, color, contrast, brightness, blur and simplicity interpreted as distribution of edges; and the Naïve Bayes and SVM for classification. Results show that high-level attributes tend to perform better than the baseline. However, the combination of the two was able to achieve even better results.

Other approaches focused on subcategories of interestingness. For instance, the authors in [27] determined "social interestingness" based on social media ranking and "visual interestingness" via crowdsourcing. The Pearson correlation coefficient between these two subcategories had low values, *e.g.,* -0.015 to 0.195, indicating that there is a difference between what people share on social networks and what has a high pure visual interest. The features used for predicting these concepts were color descriptors determined on the HSV color space, texture information via Local Binary Patterns, saliency [25] and edge information captured with Histogram of Oriented Gradients.

Individual frame interestingness was calculated by the authors in [44]. They used web photo collections of interesting landmarks from Flickr as estimators of human interest. The proposed approach involved calculating a similarity measure between each frame from YouTube travel videos and the Flickr image collection of the landmarks presented in the videos, used as interesting examples. SIFT features were computed and the number of features shared between the frame and the image collection baseline, and their spatial arrangement similarity were the components that determined the interestingness measure. Finally the authors showed that their algorithm achieved the desired results, tending to classify full images of the landmarks as interesting.

Another interesting approach is the one proposed in [32]. Authors used audio, video and high-level features for predicting video shot interestingness, *e.g.,* color histograms, SIFT [46], HOG [15, 68], SSIM Self-Similarities [56], GIST [51], MFCC [64], Spectrogram SIFT [35], Audio-Six, Classemes [65], ObjectBank [42] and the 14 photographic styles described in [49]. The system was trained via Joachims' Ranking SVM [33]. The final results showed that audio and visual features performed well, and that their fusion performed even better on the two user-annotated datasets used, giving a final accuracy of 78.6% on the 1,200 Flickr videos and 71.7% on the 420 YouTube videos. Fusion with the high-level attributes provided a better result only on the Flickr dataset, with an overall precision of 79.7% and 71.4%.

Low- and high-level features were used in [22] to detect the most interesting frames in image sequences. The selected low-level features were: raw pixel values, color histogram, HOG, GIST and image self-similarity. The high-level features were grouped in several categories: emotion predicted from raw pixel values [66], complexity defined as the size of the compressed PNG image, novelty computed through a Local Outlier Factor [8] and a learning feature computed using a ν-SVR classifier with RBF kernel on the GIST features. Each one of these features performed above the baseline (*i.e.,* random selection), and their combination also showed improvements over each individual one. The tests were performed on a database consisting

of 20 image sequences, each containing 159 color images taken from various webcams and surveillance scenarios, and the final results for the combination of features gave an average precision score of 0.35 and a $Top3$ score of 0.59.

2.3 Datasets for predicting interestingness

A critical point to build and evaluate any machine learning system is the availability of labeled data. Although the literature for automatic interestingness prediction is still at its early stages, there are some attempts to construct an evaluation data. In the following, we introduce the most relevant initiatives.

Many of the authors have chosen to create their own datasets for evaluating their methods. Various sources of information were used, mainly coming form social media, *e.g.,* Flickr [19, 27, 44, 62, 32], Pinterest [27], Youtube [44, 32]. The data consisted of the results returned by search queries. Annotations were determined either automatically, by exploiting the available social media metadata and statistics such as Flickr's "*interestingness measure*" in [19, 32], or manually, via crowdsourcing in [27] or local human assessors in [32].

The authors in [19] used a dataset composed of 40,000 images, and kept the top 10%, ordered according to the Flickr interestingness score, as positive interesting examples and the last 10% as negative, non interesting examples. Half of this dataset was used for training and half for testing. The same top and last 10% of Flickr results was used in [32], generating 1,200 videos retrieved with 15 keyword queries, *e.g.,*: "*basketball*", "*beach*", "*bird*", "*birthday*", "*cat*", "*dancing*". In addition to these, the authors in [32] also used 30 YouTube advertisement videos from 14 categories, such as "*accessories*", "*clothing&shoes*", "*computer&website*", "*digital products*", "*drink*". The videos had an average duration of 36 seconds and were annotated by human assessors, thus generating a baseline interestingness score.

Apart from the individual datasets, there were also initiatives of grouping several datasets of different compositions. The authors in [23], associated an internal context to the data: a strong context dataset proposed in [22], where the images in 20 publicly available webcam streams are consistently related to one another, thus generating a collection of 20 image sequences each containing 159 images; a weak context dataset introduced in [51] which consists of 2,688 fixed size images grouped in 8 scene categories: "*coast*", "*mountain*", "*forest*", "*open country*", "*street*", "*inside city*", "*tall buildings*" and "*highways*"; and a no context dataset which consists of the 2,222 image memorability dataset proposed in [30, 29], with no context or story behind the pictures.

3 The Predicting Media Interestingness Task

This section describes the Predicting Media Interestingness Task, which was proposed in the context of the 2016 MediaEval international evaluation campaign. This section addresses the task definition (Section 3.1), the description of the provided data with its annotations (Section 3.2), and the evaluation protocol (Section 3.3).

3.1 Task definition

Interestingness of media content is a perceptual and highly semantic notion that remains very subjective and dependent on the user and the context. Nevertheless, experiments show that there is, in general, an average and common interestingness level, shared by most of the users [10]. This average interestingness level provides evidence to envision that the building of a model for the prediction of interestingness is feasible. Starting from this basic assumption, and constraining the concept to a clearly defined use case, will serve to disambiguate the notion and reduce the level of subjectivity.

In the proposed benchmark, interestingness is assessed according to a practical use case originated from Technicolor, where the goal is to help professionals to illustrate a Video on Demand (VOD) web site by selecting some interesting frames and/or video excerpts for the movies. We adopt the following definition of interestingness: *an image/video excerpt is interesting in the context of helping a user to make his/her decision about whether he/she is interested in watching the movie it represents*. The proposed data is naturally adapted to this specific scenario, and consists of professional content, *i.e.,* Hollywood-like movies.

Given this data and use case, the task requires participants to develop systems which can automatically select images and/or video segments which are considered to be the most interesting according to the aforementioned definition. Interestingness of the media is to be judged by the systems based on visual appearance, audio information and text accompanying the data. Therefore, the challenge is inherently multi-modal.

As presented in numerous studies in the literature, predicting the interestingness level of images and videos often requires significantly different perspectives. Images are self contained and the information is captured in the scene composition and colors, whereas, videos are lower quality images in motion, whose purpose is to transmit the action via the movement of the objects. Therefore, to address the two cases, two benchmarking scenarios (subtasks) are proposed as:

- *predicting image interestingness*: given a set of key-frames extracted from a movie, the systems are required to automatically identify those images for the given movie that viewers report to be the most interesting in the given movie. To solve the task, participants can make use of visual content as well as external metadata, *e.g.,* Internet data about the movie, social media information, etc;

- *predicting video interestingness*: given the video shots of a movie, the systems are required to automatically identify those shots that viewers report to be the most interesting in the given movie. To solve the task, participants can make use of visual and audio data as well as external data, *e.g.,* subtitles, Internet data, etc.

A special feature of the provided data is the fact that it is extracted from the same source movies, *i.e.,* the key-frames are extracted from the provided video shots of the movies. Therefore, this will allow for comparison between the two tasks, namely to assess to which extent image and video interestingness are linked.

Furthermore, we proposed a binary scenario, where data can be either interesting or not (two cases). Nevertheless, a confidence score is also required for each decision, so that the final evaluation measure could be computed in a ranking fashion. This is more closely related to a real world usage scenario, where results are provided in order of decreasing interestingness level.

3.2 Data description

As mentioned in the previous section, the video and image subtasks are based on a common dataset, which consists of Creative Commons trailers of Hollywood-like movies, so as to allow redistribution. The dataset, its annotations, and accompanying features, as described in the following subsections, are publicly available[4].

The use of trailers, instead of full movies, has several motivations. Firstly, it is the need for having content that can be freely and publicly distributed, as opposed to *e.g.,* full movies which have much stronger restrictions on distribution. Basically, each copyrighted movie would require an individual permission for distribution. Secondly, using full movies is not practically feasible for the highly demanding segmentation and annotations steps with limited time and resources, as the number of images/video excerpts to process is enormous, in the order of millions. Finally, running on full movies, even if the aforementioned problems were solved, will not allow for having a high diversification of the content, as only a few movies could have been used. Trailers, will allow for selecting a larger number of movies and thus diversifying the content.

Trailers are by definition representative of the main content and quality of the full movies. However, it is important to note that trailers are already the result of some manual filtering of the movie to find the most interesting scenes, but without spoiling the movie key elements. In practice, most trailers also contain less interesting, or slower paced shots to balance their content. We therefore believe that this is a good compromise for the practicality of the data/task.

The proposed dataset is split into *development data*, intended for designing and training the algorithms which is based on 52 trailers; and *testing data* which is used for the actual evaluation of the systems, and is based on 26 trailers.

---

[4] http://www.technicolor.com/en/innovation/scientific-community/scientific-data-sharing/interestingness-dataset

The data for the video subtask was created by segmenting the trailers into video shots. The same video shots were also used for the image subtask, but here each shot is represented by a single key-frame image. The task is thus to classify the shots, or key-frames, of a particular trailer, into interesting and non interesting samples.

3.2.1 Shot segmentation and key-frame extraction

Video shot segmentation was carried out manually using a custom-made software tool. Here we define a video shot as a continuous video sequence recorded between a turn-on and a turn-off of the camera. For an edited video sequence, a shot is delimited between two video transitions. Typical video transitions include sharp transitions or cuts (direct concatenation of two shots), and gradual transitions like fades (gradual disappearance/appearance of a frame to/from a black frame) and dissolves (gradual transformation of one frame into another). In the process, we discarded movie credits and title shots. Gradual transitions were considered presumably very uninteresting shots by themselves, whenever possible. In a few cases, shots in between two gradual transitions were too short to be segmented. In that case, they were merged with their surrounding transitions, resulting in one single shot.

The segmentation process resulted in 5,054 shots for the development dataset, and 2,342 shots for the test dataset, with an average duration of one second in each case. These shots were used for the video subtask. For the image subtask, we extracted a single key-frame for each shot. The key-frame was chosen as the middle frame, as it is likely to capture the most representative information of the shot.

3.2.2 Ground-truth annotation

All video shots and key-frames were manually annotated in terms of interestingness by human assessors. The annotation process was performed separately for the video and image subtasks, to allow us to study the correlation between the two. Indeed we would like to answer the question: Does image interestingness automatically imply video interestingness, and vice versa?

A dedicated web-based tool was developed to assist the annotation process. The tool has been released as free and open source software, so that others can benefit from it and contribute improvements[5].

We use the following annotation protocol. Instead of asking annotators to assign an interestingness value to each shot/key-frame, we used a pair-wise comparison protocol where the annotators were asked to select the more interesting shot/key-frame from a pair of examples taken from the same trailer. Annotators were provided with the clips for the shots and the images for the key-frames, presented side by side. Also, they were informed about the Video on Demand-use case, and asked to consider also that “the selected video excerpts/key-frames should be suitable in

[5] https://github.com/mvsjober/pair-annotate

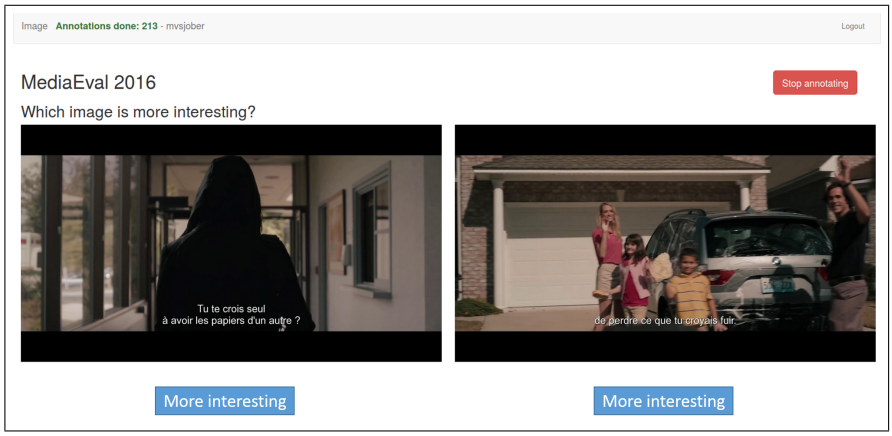


Fig. 1: Web user interface for pair-wise annotations.

terms of helping a user to make his/her decision about whether he/she is interested in watching a movie". Figure 1 illustrates the pair-wise decision stage of the user interface.

The choice of a pair-wise annotation protocol instead of direct rating was based on our previous experience with annotating multimedia for affective content and interestingness [61, 3, 10]. Assigning a rating is a cognitively very demanding task, requiring the annotator to understand, and constantly keep in mind, the full range of the interestingness scale [70]. Making a single comparison is a much easier task as one only needs to compare the interestingness of two items, and not consider the full range. Directly assigning a rating value is also problematic since different annotators may use different ranges, and even for the same annotator the values may not be easily interpreted [52]. For example, is an increase from 0.3 to 0.4 the same as the one from 0.8 to 0.9? Finally, it has been shown that pairwise comparisons are less influenced by the order in which the annotations are displayed than with direct rating [71].

However, annotating all possible pairs is not feasible due to the sheer number of comparisons required. For instance, n shots/key-frames would require $n(n-1)/2$ comparisons to be made for a full coverage. Instead, we adopted the adaptive square design method [41], where the shots/key-frames are placed in a square design and only pairs on the same row or column are compared. This reduces the numbers of comparisons to $n(\sqrt{n}-1)$. For example, for $n = 100$ we need to undergo only 900 comparisons instead of $4,950$ (full coverage). Finally, the Bradley-Terry-Luce (BTL) model [7] was used to convert the paired comparison data to a scalar value.

We modified the adaptive square design setup so that comparisons were taken by many users simultaneously until all the required pairs had been covered. For the rest, we proceeded according to the scheme in [41]:

1. Initialization: shots/key-frames are randomly assigned positions in the square matrix;
2. Perform a single annotation round according to the shot/key-frame pairs given by the square (across rows, columns);
3. Calculate the BTL scores based on the annotations;
4. Re-arrange the square matrix so that shots/key-frames are ranked according to their BTL scores, and placed in a spiral. This arrangement ensures that mostly similar shots/key-frames are compared row-wise and column-wise;
5. Repeat steps 2. to 4. until convergence.

For practical reasons, we decided to consider by default that convergence is achieved after 5 rounds and thus terminated the process when the five runs are finished. The final binary interestingness decisions were obtained with a heuristic method that tried to detect a "jumping point" in the normalized distribution of the BTL values for each movie separately. The underlying motivation for this empirical rule is the assumption that the distribution is a sum of two underlying distributions: non interesting shots/key-frames, and interesting shots/key-frames.

Overall, 315 annotators participated in the annotation for the video data and 100 for the images. The cultural distribution is over 29 different countries around the world. The average reported age of the annotators was 32, with a standard deviation around 13. Roughly, 66% were male, 32% female, and 2% did not specify their gender.

3.2.3 Additional features

Apart from the data and its annotations, to broaden the targeted communities, we also provide some pre-computed content descriptors, namely:

Dense SIFT which are computed following the original work in [46], except that the local frame patches are densely sampled instead of using interest point detectors. A codebook of 300 codewords is used in the quantization process with a spatial pyramid of three layers [40].

HoG descriptors *i.e.,* Histograms of Oriented Gradients [15] are computed over densely sampled patches. Following [68], HoG descriptors in a 2×2 neighborhood are concatenated to form a descriptor of higher dimension.

LBP *i.e.,* Local Binary Patterns as proposed in [50].

GIST is computed based on the output energy of several Gabor-like filters (8 orientations and 4 scales) over a dense frame grid like in [51].

Color histogram computed in the HSV space (Hue-Saturation-Value).

MFCC computed over 32ms time-windows with 50% overlap. The cepstral vectors are concatenated with their first and second derivatives.

CNN features *i.e.,* the *fc7 layer* (4,096 dimensions) and *prob layer* (1,000 dimensions) of AlexNet [31].

Mid level face detection and tracking features obtained by face tracking-by-detection in each video shot via a HoG detector [15] and the correlation tracker proposed in [16].

3.3 Evaluation Rules

As for other tasks in MediaEval, participants were allowed to submit a total of up to 5 runs for the video and image subtasks. To provide the reader with a complete picture of the evaluation process in order to understand the achieved results, we replicate the exact conditions for the participants, here.

Each task had a required run, namely: for predicting image interestingness, classification had to be achieved with the use of the visual information only, no external data was allowed; for predicting video interestingness, classification had to be achieved with the use of both audio and visual information; no external data was allowed. External data was considered to be any of the following: additional datasets and annotations which were specifically designed for interestingness classification; the use of pre-trained models, features, detectors obtained from such dedicated additional datasets; additional metadata from the Internet (*e.g.,* from IMDb). On the contrary, CNN features trained on generic datasets such as ImageNet were allowed for use in the required runs. By generic datasets, we mean datasets that were not explicitly designed to support research in interestingness prediction. Additionally, datasets dedicated to study memorability or other aspects of media were allowed, as long as these concepts are different from interestingness, although a correlation may exist.

To assess performance, several metrics were computed. The official evaluation metric was the mean average precision (MAP) computed over all trailers, whereas average precision was to be computed on a per trailer basis, over all ranked images/video shots. MAP was computed with the trec_eval tool[6]. In addition to MAP, several other secondary metrics were provided, namely: accuracy, precision, recall and f-score for each class, and the class confusion matrix.

4 Results and analysis of the first benchmark

4.1 Official results

The 2016 Predicting Media Interestingness Task received more than 30 registrations and 12 teams coming from 9 countries all over the world submitted runs in the end (see Figure 2). The task attracted a lot of interest from the community, which shows the importance of this topic.

Tables 1 and 2 provide an overview of the official results for the two subtasks (video and image interestingness prediction). A total of 54 runs were received, equally distributed between the two subtasks. As a general conclusion, the achieved MAP values were low, which proves again the challenging nature of this problem. Slightly higher values were obtained for image interestingness prediction.

[6] http://trec.nist.gov/trec_eval/

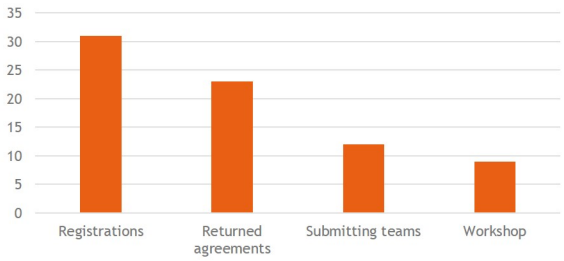


Fig. 2: 2016 Predicting Media Interestingness task's participation at different stages.

To serve as a baseline for comparison, we generated a random ranking run, *i.e.,* samples were ranked randomly 5 times and we take the average MAP. Compared to the baseline, the results of the image subtask clearly confirm their performance, being almost all above the baseline. For the video subtask, on the other hand, the value range is smaller and a few systems did worse than the baseline. In the following we present the participating systems and analyze the achieved results in detail.

4.2 Participating systems and global trends

Numerous approaches have been investigated by the participating teams to tackle both image and video interestingness prediction. In the following, we will firstly summarize the general techniques used by the teams and their key features (Section 4.2.1), and secondly present the global insights of the results (Section 4.2.2).

4.2.1 Participants' approaches

A summary of the features and classification techniques used by each participating system is presented in Table 3 (image interestingness) and Table 4 (video interestingness). Below, we present the main characteristics of each approach. Unless otherwise specified, each team participated in both subtasks.

BigVid [69] (Fudan University, China): explored various low-level features (from visual and audio modalities) and high-level semantic attributes, as well as the fusion of these features for classification. Both SVM and recent deep learning methods were tested as classifiers. The results proved that the high-level attributes are complementary to visual features since the combination of these features increases the overall performance.

ETH-CVL [67] (ETH Zurich, Switzerland): participated in the video subtask only. Two models were presented: (i) a frame-based model that uses textual side information (external data) and (ii) a generic predictor for finding video highlights in the form of segments. For the frame-based model, they learned a joint embed-

Team	Run name	MAP
TUD-MMC [43]	me16in_tudmmc2_image_histface	0.2336
Technicolor [57]	me16in_technicolor_image_run1_SVM_rbf	0.2336
Technicolor	me16in_technicolor_image_run2_DNNresampling06_100	0.2315
MLPBOON [53]	me16in_MLPBOON_image_run5	0.2296
BigVid [69]	me16in_BigVid_image_run5FusionCNN	0.2294
MLPBOON	me16in_MLPBOON_image_run1	0.2205
TUD-MMC	me16in_tudmmc2_image_hist	0.2202
MLPBOON	me16in_MLPBOON_image_run4	0.217
HUCVL [21]	me16in_HUCVL_image_run1	0.2125
HUCVL	me16in_HUCVL_image_run2	0.2121
UIT-NII [39]	me16in_UITNII_image_FA	0.2115
RUC [12]	me16in_RUC_image_run2	0.2035
MLPBOON	me16in_MLPBOON_image_run2	0.2023
HUCVL	me16in_HUCVL_image_run3	0.2001
RUC	me16in_RUC_image_run3	0.1991
RUC	me16in_RUC_image_run1	0.1987
ETH-CVL [67]	me16in_ethcvl1_image_run2	0.1952
MLPBOON	me16in_MLPBOON_image_run3	0.1941
HKBU [45]	me16in_HKBU_image_baseline	0.1868
ETH-CVL	me16in_ethcvl1_image_run1	0.1866
ETH-CVL	me16in_ethcvl1_image_run3	0.1858
HKBU	me16in_HKBU_image_drbaseline	0.1839
BigVId	me16in_BigVid_image_run4SVM	0.1789
UIT-NII	me16in_UITNII_image_V1	0.1773
LAPI [14]	me16in_lapi_image_runf1	0.1714
UNIGECISA [54]	me16in_UNIGECISA_image_ReglineLoF	0.1704
baseline		0.16556
LAPI	me16in_lapi_image_runf2	0.1398

Table 1: Official results for image interestingness prediction evaluated by MAP.

ding space for image and text, which allows to measure relevance of a frame with regard to some text such as the video title. For video interestingness prediction, the approach in [24] was used, where a deep RankNet is trained to rank the segments of a video based upon their suitability as animated GIFs. Note that RankNet captures the spatio-temporal aspect of video segments via the use of 3D convolutional neural networks (C3D).

HKBU [45] (Hong Kong Baptist University, China): used two dimensionality reduction methods, named Neighborhood MinMax Projections (NMMP) and Supervised Manifold Regression (SMR), to extract features of lower dimension from a set of baseline low-level visual features (Color Histogram, dense SIFT, GIST, HOG, LBP). Then nearest neighbor (NN) classifier and Support Vector Regressor (SVR) were exploited for interestingness classification. They found that after dimensionality reduction, the performance of the reduced features was comparable to that of their original features, which indicated that the reduced features successfully captured most of the discriminant information of the data.

Team	Run name	MAP
UNIFESP [1]	me16in_unifesp_video_run1	0.1815
HKBU [45]	me16in_HKBU_video_drbaseline	0.1735
UNIGECISA [54]	me16in_UNIGECISA_video_RegsrrLoF	0.171
RUC [12]	me16in_RUC_video_run2	0.1704
UIT-NII [39]	me16in_UITNII_video_A1	0.169
UNIFESP	me16in_unifesp_video_run4	0.1656
RUC	me16in_RUC_video_run1	0.1647
UIT-NII	me16in_UITNII_video_F1	0.1641
LAPI [14]	me16in_lapi_video_runf5	0.1629
Technicolor [57]	me16in_technicolor_video_run5_CSP_multimodal_80_epoch7	0.1618
UNIFESP	me16in_unifesp_video_run2	0.1617
UNIFESP	me16in_unifesp_video_run3	0.1617
ETH-CVL [67]	me16in_ethcvl1_video_run2	0.1574
LAPI	me16in_lapi_video_runf3	0.1574
LAPI	me16in_lapi_video_runf4	0.1572
TUD-MMC [43]	me16in_tudmmc2_video_histface	0.1558
TUD-MMC	me16in_tudmmc2_video_hist	0.1557
BigVid [69]	me16in_BigVid_video_run3RankSVM	0.154
HKBU	me16in_HKBU_video_baseline	0.1521
BigVid	me16in_BigVid_video_run2FusionCNN	0.1511
UNIGECISA	me16in_UNIGECISA_video_RegsrrGiFe	0.1497
baseline		0.1496
BigVid	me16in_BigVid_video_run1SVM	0.1482
Technicolor	me16in_technicolor_video_run3_LSTM_U19_100_epoch5	0.1465
UNIFESP	me16in_unifesp_video_run5	0.1435
UNIGECISA	me16in_UNIGECISA_video_SVRloAudio	0.1367
Technicolor	me16in_technicolor_video_run4_CSP_video_80_epoch9	0.1365
ETH-CVL	me16in_ethcvl1_video_run1	0.1362

Table 2: Official results for video interestingness prediction evaluated by MAP.

HUCVL [21] (Hacettepe University, Turkey): participated in image interestingness prediction only. They investigated three different Deep Neural Network (DNN) models. The first two models were based on fine-tuning two pre-trained models, namely AlexNet and MemNet. Note that MemNet was trained on the image memorability dataset proposed in [37], the idea being to see if memorability can be generalized to the interestingness concept. The third model, on the other hand, depends on a proposed triplet network which comprised three instances with shared weights of the same feed-forward network. The results demonstrated that all these models provide relatively similar and promising results on the image interestingness subtask.

LAPI [14] (University Politehnica of Bucharest, Romania, co-organizer of the task): investigated a classic descriptor-classification scheme, namely the combination of different low-level features (HoG, dense SIFT, LBP, GIST, AlexNet fc7 layer features (hereafter referred as CNN features), Color Histogram, Color Naming Histogram) and use of SVM, with different kernel types, as classifier. For video, frame features were averaged to obtain a global video descriptor.

Team	Features	Classification technique
BigVid [69]	denseSIFT+CNN+Style Attributes+SentiBank	SVM (run4) Regularized DNN (run5)
ETH-CVL [67]	DNN-based	Visual Semantic Embedding Model
HKBU [45]	ColorHist+denseSIFT+GIST+HOG+LBP (run1) features from run1 + dimension reduction (run2)	Nearest Neighbor and SVR
HUCVL [21]	CNN (run1, run3) MemNet (run2)	MLP (run1, run2) Deep triplet network (run3)
LAPI [14]	ColorHist+GIST (run1) denseSIFT+GIST (run2)	SVM
MLPBOON [53]	CNN, PCA for dimension reduction	Logistic Regression
RUC [12]	GIST+LBP+CNN prob (run1) ColorHist+GIST+CNN prob (run2), ColorHist+GIST+LBP+CNN prob (run3)	Random Forest (run1) Random Forest (run2) SVM (run3)
Technicolor [57]	CNN (Alexnet fc7)	SVM (run1) MLP (run2)
TUD-MMC [43]	Face-related ColorHist (run1) Face-related ColorHist+Face area (run2)	Normalized histogram-based confidence score (NHCS) (run1) NHCS+Normalized face area score (run2)
UIT-NII [39]	CNN (AlexNet+VGG) (run1) CNN (VGG)+GIST+HOG+DenseSIFT (run2)	SVM with late fusion
UNIGECISA [1]	Multilingual visual sentiment ontology (MVSO)+CNN	Linear Regression

Table 3: Overview of the characteristics of the submitted systems for predicting image interestingness.

MLPBOON [53] (Indian Institute of Technology, Bombay, India): participated only in image interestingness prediction and studied various baseline visual features provided by the organizers [18], and classifiers on the development dataset. Principal component analysis (PCA) was used for reducing the feature dimension. Their final system involved the use of PCA on CNN features for the input representation and logistic regression (LR) as classifier. Interestingly, they observed that the combination of CNN features with GIST and Color Histogram features gave similar performance to the use of CNN features only. Overall, this simple, yet effective, system obtained quite high MAP values for the image subtask.

RUC [12] (Renmin University, China): investigated the use of CNN features and AlexNet probabilistic layer (referred as CNN prob), and hand-crafted visual features including Color Histogram, GIST, LBP, HOG, dense SIFT. Classifiers were SVM and Random Forest. They found that semantic-level features, *i.e.,* CNN prob, and low-level appearance features are complementary. However, concatenating CNN features with hand-crafted features did not bring any improvement. This finding is coherent with the statement from MLPBOON team [53]. For predicting video interestingness, audio modality offered superior performance than visual modality and the early fusion of the two modalities can further boost the performance.

Technicolor [57] (Technicolor R&D France, co-organizer of the task): used CNN features as visual features (for both the image and video subtasks), and MFCC as au-

Teams	Features	Classification technique	Multi-modality
BigVid [69]	denseSIFT, CNN Stype Attrubutes, SentiBank	SVM (run1) Regularlized DNN (run2) SVM/Ranking-SVM (run3)	No
ETH-CVL [67]	DNN-based	Video2GIF (run1) Video2GIF+Visual Semantic Embedding Model (run2)	Text+Visual
HKBU [45]	ColorHist+denseSIFT+GIST +HOG+LBP (run1) features from run1 + dimension reduction (run2)	Nearest Neighbor and SVR	No
LAPI [14]	GIST+CNN prob (run3) ColorHist+CNN (run4) denseSIFT+CNN prob (run5)	SVM	No
RUC [12]	Acoustic Statistics + GIST (run4) MFCC with Fisher Vector Encoding + GIST (run5)	SVM	Audio+Visual
Technicolor [57]	CNN+MFCC	LSTM-Resnet + MLP (run3) Proposed RNN-based model (run4, run5)	Audio+Visual
TUD-MMC [43]	ColorHist (run1) ColorHist+Face area (run2)	Normalized histogram-based confidence score (NHCS) run3) NHCS+Normalized face area score (run4)	No
UIT-NII [39]	CNN (AlexNet)+MFCC (run3) CNN (VGG)+GIST (run4)	SVM with late fusion	Audio+Visual
UNIFESP [1]	Histogram of motion patterns (HMP) [2]	Majority Voting of pairwise ranking methods: Ranking SVM, RankNet RankBoost, ListNet	No
UNIGECISA [54]	MVSO+CNN (run2) Baseline visual features [18] (run3), Emotionally-motivated audio feature (run4)	SVR (run2) SPARROW (run3, run4)	Audio+Visual

Table 4: Overview of the characteristics of the submitted systems for predicting video interestingness.

dio feature (for the video subtask) and investigated the use of both SVM and different Deep Neural Networks (DNN) as classification techniques. For the image subtask, a simple system with CNN features and SVM resulted in the best MAP, 0.2336. For the video subtask, multi-modality as a mid-level fusion of audio and visual features, was taken into account within the DNN framework. Additionally, a novel DNN architecture based on multiple Recurrent Neural Networks (RNN) was proposed for modeling the temporal aspect of the video, and a resampling/upsampling technique was used to deal with the unbalanced dataset.

TUD-MMC [43] (Delft University of Technology, Netherlands): investigated MAP values obtained on the development set by swapping and submitting ground-truth annotations of image and video to the video and image subtasks respectively,

i.e., using the video ground-truth as submission on the image subtask and the image ground-truth as submission on the video subtask. They concluded on the low correlation between the image interestingness and video interestingness concepts. Their simple visual features took into account the human face information (color and sizes) in the image and video with the assumption that clear human faces should attract the viewers attention and thus make the image/video more interesting. One of their submitted runs, only rule-based, obtained the best MAP value of 0.2336 for the image subtask.

UIT-NII [39] (University of Science, Vietnam; University of Information Technology, Vietnam; National Institute of Informatics, Japan): used SVM to predict three different scores given the three types of input features: (1) low-level visual features provided by the organizers [18], (2) CNN features (AlexNet and VGG), and (3) MFCC as audio feature. Late fusion of these scores was used for computing the final interestingness levels. Interestingly, their system tends to output a higher rank on images of beautiful women. Furthermore, they found that images from dark scenes were often considered as more interesting.

UNIFESP [1] (Federal University of Sao Paulo, Brazil): participated only in the video subtask. Their approach was based on combining learning-to-rank algorithms for predicting the interestingness of videos by using their visual content only. For this purpose, Histogram of Motion Patterns (HMP) [2] were used. A simple majority voting scheme was used for combining 4 pairwise machine learned rankers (Ranking SVM, RankNet, RankBoost, ListNet) and predicting the interestingness of videos. This simple, yet effective, method obtained the best MAP of 0.1815 for the video subtask.

UNIGECISA [54] (University of Geneva, Switzerland): used mid-level semantic visual sentiment features, which are related to the emotional content of images and were shown to be effective in recognizing interestingness in GIFs [24]. They found that these features outperform the baseline low-level ones provided by the organizers [18]. They also investigated the use of emotionally-motivated audio features (eGeMAPS) for the video subtask and showed the significance of the audio modality. Three regression models were reported to predict interestingness levels: linear regression (LR), SVR with linear kernel, and sparse approximation weighted regression (SPARROW).

4.2.2 Analysis of this year's trends and outputs

This section provides an in-depth analysis of the results and discusses the global trends found in the submitted systems.

Low-level vs. high-level description: The conventional low-level visual features, such as dense SIFT, GIST, LBP, Color Histogram, were still being used by many of the systems for both, image and video interestingness prediction [14, 45, 39, 69, 12]. However, deep features like CNN features (*i.e.,* Alexnet fc7 or VGG) have become dominant and are exploited by the majority of the systems. This shows the effec-

tiveness and popularity of deep learning. Some teams investigated the combination of hand crafted features with deep features, *i.e.,* conventional and CNN features. A general finding is that such a combination did not really bring any benefit to the prediction results [45, 12, 53]. Some systems combined low-level features with some high-level attributes such as emotional expressions, human faces, CNN visual concept predictions [69, 12]. In this case, the resulting conclusion was that low-level appearance features and semantic-level features are complementary, as the combination in general offered better prediction results.

Standard vs. deep learning-based classification: As it can be seen in Tables 3 and 4, SVM was mostly used by a large number of systems, for both prediction tasks. In addition, regression techniques such as linear regression, logistic regression, and support vector regression were also widely reported. Contrary to CNN features, which were widely used by most of the systems, deep learning classification techniques were investigated less (see [57, 67, 21, 69] for image interestingness and [57, 67, 69] for the video interestingness). This may be due to the fact that the datasets are not large enough to justify a deep learning approach. Conventional classifiers were preferred here.

Use of external data: Some systems investigated the use of external data to improve the results. For instance, Flickr images with social-driven interestingness labels were used for model selection in the image interestingness subtask by the Technicolor team [57]. The HUCVL team [21] submitted a run with a fine-tuning of the MemNet model, which was trained for image memorability prediction. Although memorability and interestingness are not the same concept, the authors expected that fine-tuning a model related to an intrinsic property of images could be helpful in learning better high-level features for image interestingness prediction. The ETH-CVL team [67] exploited movie titles, as textual side information related to movies, for both subtasks. In addition, ETH-CVL also investigated the use of the deep RankNet model, which was trained on the Video2GIF dataset [24], and the Visual Semantic Embedding model, which was trained on the MSR Clickture dataset [28].

Dealing with small and unbalanced data: As the development data provided for the two subtasks are not very large, some systems, *e.g.,* [57, 1], used the whole image and video development sets for training when building the final models. To cope with the imbalance of the two classes in the dataset, the Technicolor team [57] proposed to use classic resampling and upsampling strategies so that the positive samples are used multiple times during training.

Multi-modality: Specific to video interestingness, multi-modal approaches were exploited by half of the teams for at least one of their runs, as shown in Table 4. Four teams combined audio and visual information [57, 12, 39, 54], and one team combined text with visual information [67]. The fusion of modalities was done either at the early stage [12, 54], middle stage [57], or late stage [39] in the processing workflows. Note that the combination of text and visual information was also reported in [67] for image interestingness prediction. The general finding here was that multi-modality brings benefits to the prediction results.

Temporal modeling for video: Though the temporal aspect is an important property of a video, most systems did not actually exploit any temporal modeling for video interestingness prediction. They mainly considered a video as a sequence of frames and a global video descriptor was computed simply by averaging frame image descriptors over each shot. As an example, HKBU team [45] treated each frame as a separated image, and calculated the average and standard deviation of their features over all frames in a shot to build their global feature vector for each video. Only two teams incorporated temporal modeling in their submitted systems, namely Technicolor [57] who used long-short term memory (LSTM) in their deep learning-based framework, and ETH-CVL [67] who used 3D convolutional neural networks (C3D) in their video highlight detector, trained on the Video2GIF dataset.

4.3 In-depth analysis of the data and annotations

The purpose of this section is to give some insights on the characteristics of the produced data, *i.e.,* the dataset and its annotations.

4.3.1 Quality of the dataset

In general, the overall results obtained during the 2016 campaign show low values for MAP (see Figures 1 and 2), especially for the video interestingness prediction subtask. To have a comparison, we provide examples of MAP values obtained by other multi-modal tasks from the literature. Of course, these were obtained on other datasets which are fundamentally different from the underlying data, both from the data point of view and also use case scenario. A direct comparison is not possible, however, they provide an idea about the current classification capabilities for video:

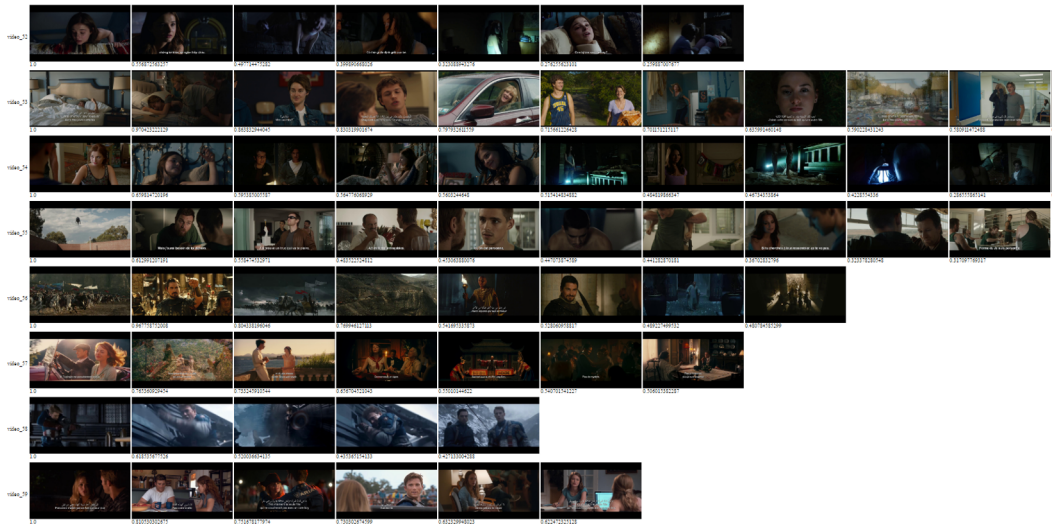
- ILSVR Challenge 2015, Object Detection with provided training data, 200 fully labeled categories, best MAP is 0.62; Object Detection from videos with provided training data, 30 fully labeled categories, best MAP is 0.67;
- TRECVID 2015, Semantic indexing of concepts such as: *airplane, kitchen, flags, etc*, best MAP is 0.37;
- TRECVID 2015, Multi-modal event detection, *e.g., somebody cooking on an outdoor grill*, best MAP is less than 0.35.

Although higher than the obtained MAP for the Predicting Media Interestingness Task, it must be noted that for more difficult tasks such as multi-modal event detection, the difference of performance is not that high, given the fact that the proposed challenge is far more subjective than the tasks we are referring to.

Nevertheless, we may wonder, especially for the video interestingness subtask, whether the quality of the dataset/anotations partly affects the predicting performance. Firstly, the dataset size, although it is sufficient for classic learning tech-

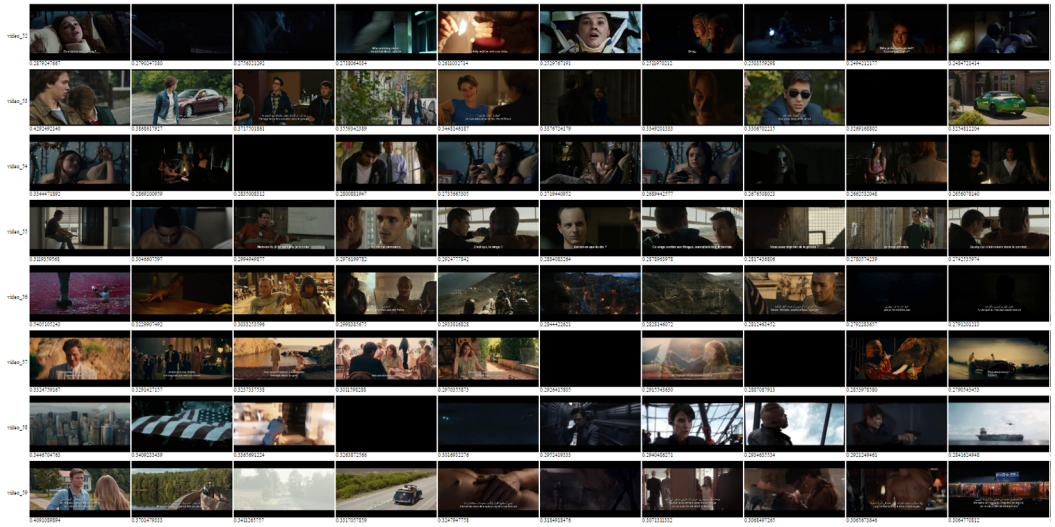
niques and required a huge annotation effort, it may not be sufficient for deep learning, with only several thousands of samples for both substasks.

Furthermore, it may be considered to be highly unbalanced with 8.3% and 9.6% of interesting content for the development set and test set, respectively. Trying to cope with the dataset's unbalance has shown to increase the performance for some systems [57, 58]. This leads to the conclusion that, although this unbalance reflects reality, *i.e.,* interesting content corresponds to only a small part of the data, it makes the task even more difficult, as systems will have to take this characteristic into account.



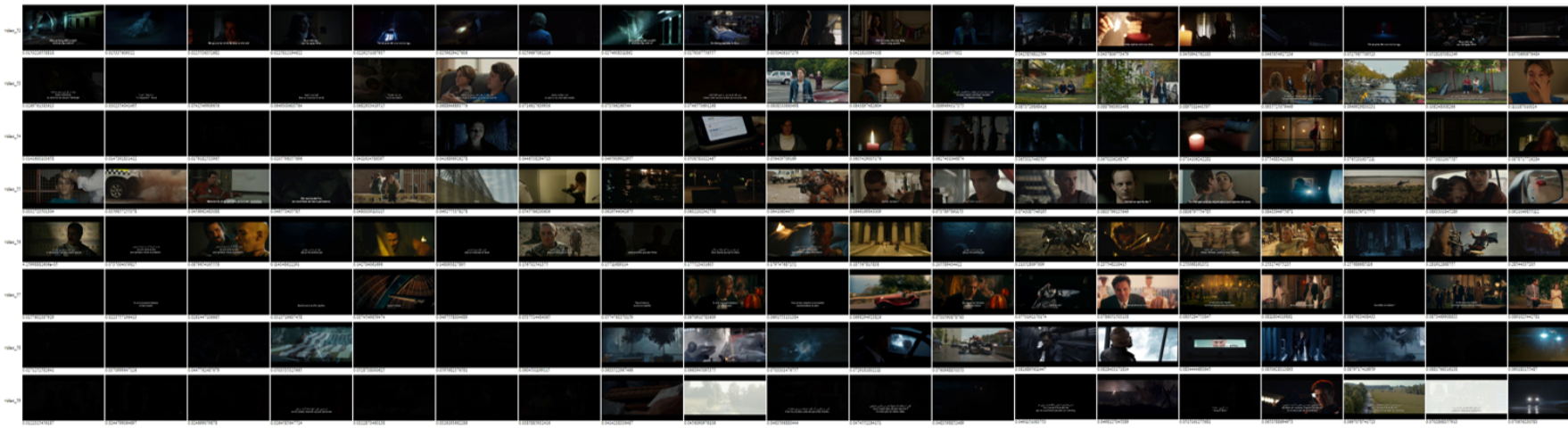
(a) Interesting images according to the ground-truth.

(b) Interesting images selected by the best system.

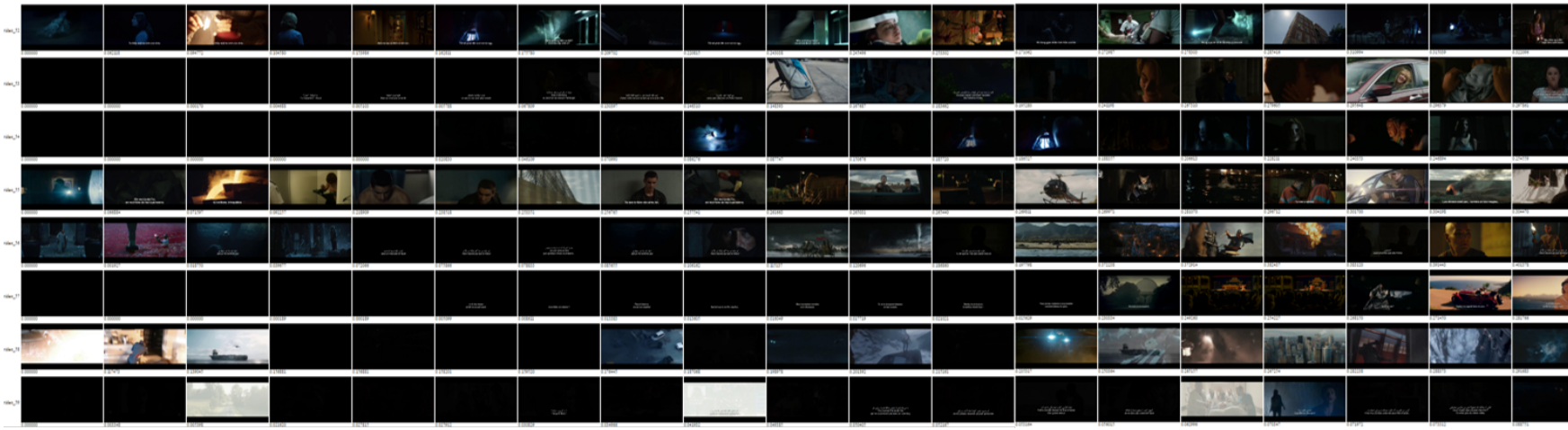


(c) Interesting images selected by the second worst performing system.

Fig. 3: Examples of interesting images from different videos of the test set. Images are ranked from left to right decreasing interestingness ranking.



(a) Non interesting images according to the ground-truth.



(b) Non interesting images selected by the best system.

Fig. 4: Examples of non interesting images from different videos of the test set. Images are ranked from left to right increasing interestingness ranking.

Finally, in Section 3.2, we explained that the final annotations were determined with an iterative process which required the convergence of the results. Due to limited time and human resources, this process was limited to 5 rounds. More rounds would certainly have resulted in better convergence of the inter-annotator ratings.

To have an idea of the subjective quality of the ground-truth rankings, Figures 3 and 4 illustrate some image examples for the image interestingness subtask together with the rankings obtained by one of the best systems and the second worst performing system, for both interesting and non interesting images.

The figures show that results obtained by the best system for the most interesting images are coherent with the selection proposed by the ground-truth, whereas the second worst performing system offers more images at the top ranks which do not really contain any information, *e.g.,* black or uniform frames, with blur or objects and persons only partially visible.

These facts converge to the idea that both the provided ground-truth and the best working systems have managed to capture the interestingness of images. It also confirms that the obtained MAP values, although quite low, nevertheless correspond to real differences in the interestingness prediction performance.

The observation of the images which were classified as non interesting (Figure 4) is also a source of interesting insights. According to the ground-truth and also to the best performing systems, non interesting images tend to be those mostly uniform, of low quality or without meaningful information. The amount of information contained in the non interesting images then increases with the level of interestingness. Note that we do not show here the images classified as non interesting by the second worst performing system, as we did for the interesting images, because there were too few (for the example 7 images out of 25 videos) to draw any conclusion.

We also calculated Krippendorff's alpha metric (α), which is a measure for inter-observer agreement [38, 26], to be $\alpha = 0.059$ for image interestingness and $\alpha = 0.063$ for video interestingness. This result would indicate that there is no inter-observer agreement. However, as our method (by design) produced very few duplicate comparisons it is not clear if this result is reliable.

As a last insight, it is worth noting that the two experienced teams [67, 54], *i.e.,* the two teams that did work on predicting content interestingness before the MediaEval benchmark, did not achieve particularly good results on both subtasks and especially on the image subtask. This raises the question of the generalization ability of their systems on different types of content, unless this difference of performance comes from the choice of different use cases as working context. For the latter, this seems to show that, to different use cases correspond different interpretations of the interestingness concept.

4.3.2 Correlation between the two subtasks

The Predicting Media Interestingness task was designed so that a comparison between the interestingness prediction for images and videos would be possible afterwards. Indeed, the same videos were used to extract both the shots and the key-

frames to be classified in each subtask, each key-frame corresponding to the middle of shots. Thanks to this, we studied a potential correlation between image interestingness and video interestingness.

Figure 5 shows the annotated video ranking against their key-frame ranking for several videos in the development set. None of the curves exhibit a correlation (the coefficient of determination, *R-squared* or R^2, used while fitting a regression line to the data, exhibits values lower than 0.03), leading to the conclusion that the two concepts differ, in the sense that we cannot use video interestingness to infer the image interestingness and the other way round on this data and use case scenario.

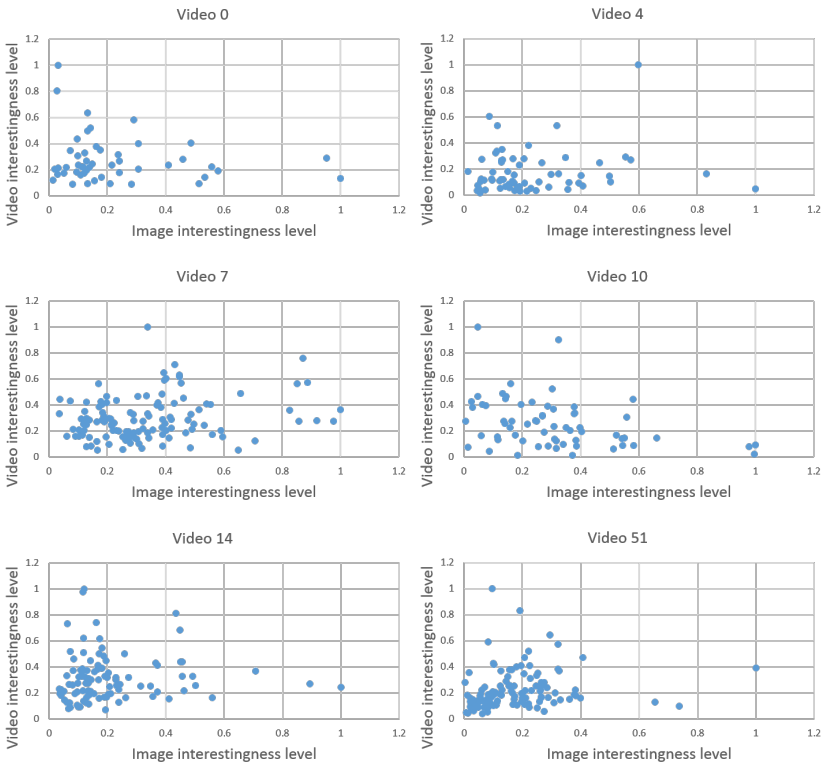


Fig. 5: Representation of image rankings vs. video rankings from the ground-truth for several videos of the development set.

This conclusion is in line with what was found in [43] where the authors investigated the assessment of the ground-truth ranking of the image substask against the ground-truth ranking of the video subtask and vice-versa. MAP value achieved by the video ground-truth on the image subtask was 0.1747, while for the image ground-truth on the video subtask, it was 0.1457, *i.e.,* in the range, or even lower, than the random baseline for both cases. Videos obviously contain more informa-

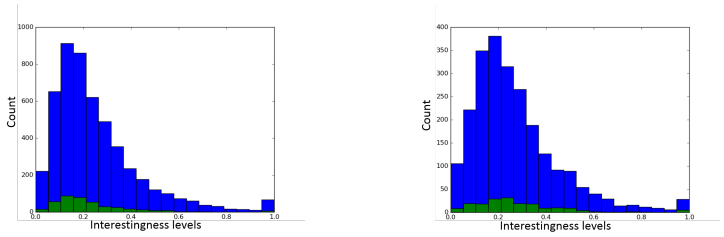
tion than a single image, which can be conveyed by other channels such as audio and motion, for example. Because of this additional information, a video might be globally considered as interesting while one single key-frame extracted from the same video will be considered as non interesting. This can explain, in some cases, the observed discrepancy between image and video interestingnesses.

4.3.3 Link with perceptual content characteristics

Trying to infer some potential links between the interestingness concept and perceptual content characteristics, we did study how low-level characteristics such as shot length, average luminance, blur and presence of high quality faces influence the interestingness prediction of images and videos.

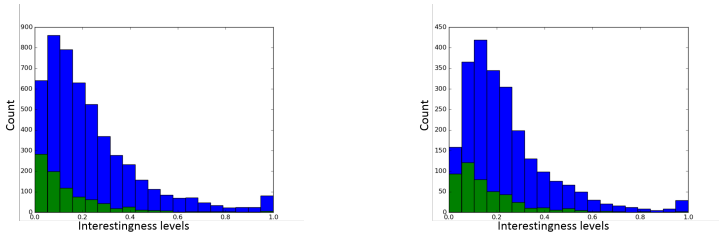
A first qualitative study of both sets of interesting and non interesting images in the development and test sets shows that most uniformly black and very blurry images were mostly classified as non interesting. So were the majority of images with no real information, *i.e.,* close-up of usual objects, partly cut faces or objects, etc., as it can be seen in Figure 4.

Figure 6 shows the distributions of interestingness values for both the development and test sets, in the video interestingness subtask, compared to the distributions of interesting values restricted to the shots with less than 10 frames. In all cases, it seems that the distributions of small shots can just be superimposed under the complete distributions, meaning that the shot length does not seem to influence the interestingness of video segments even for very short durations. On the contrary, Figure 7 shows the two same types of distributions but for the image interestingness subtask and when trying to assess the influence of the average luminance value on interestingness. This time, the distributions of interestingness levels for the images with low average luminance seem to be slightly shifted toward lower interestingness levels. This might lead us to the conclusion that low average luminance values tend to decrease the interestingness level of a given image, contrary to the conclusion in [39].



(a) Development set (b) Test set

Fig. 6: Video interestingness and shot length: distribution of interestingness levels (in blue — all shots considered; in green — shots with length smaller than 10 frames).



(a) Development set (b) Test set

Fig. 7: Image interestingness and average luminance: distribution of interestingness levels (in blue — all key-frames considered; in green — key-frames with luminance values lower than 25).

We also investigated some potential correlation between the presence of high-quality faces in frames and the interestingness level. By high-quality faces, we mean rather big faces with no motion blur, either frontal or profile, no closed eyes or funny faces. This last mid-level characteristic was assessed manually by counting the number of high-quality faces present in both the interesting and non interesting images for the image interestingness subtask. The proportion of high-quality faces on the development set was found to be 48.2% for the set of images annotated as interesting and 33.9% for the set of images annotated as non interesting. For the test set, 56.0% of the interesting images and 36.7% of the non interesting images contain high quality faces. The difference in favor of the interesting sets tends to prove that this characteristic has a positive influence on the interestingness assessment. This was confirmed by the results obtained by TUD-MMC team [43] who based their system only on the detection of these high quality faces and achieved the best MAP value for the image subtask.

As a general conclusion, we may say that perceptual quality plays an important role when assessing the interestingness of images, although it is not the only clue to assess the interestingness of content. Among other semantic objects, the presence of good quality human faces seems to be correlated with interestingness.

5 Conclusions and future challenges

In this chapter we introduced a specially designed evaluation framework for assessing the performance of automatic techniques for predicting image and video interestingness. We described the released dataset and its annotations. Content interestingness was defined in a multi-modal scenario and for a real-world, specific use case defined by Technicolor R&D France, namely the selection of interesting images and video excerpts for helping professionals to illustrate a Video on Demand (VOD) web site.

The proposed framework was validated during the 2016 Predicting Media Interestingness Task, organized with the MediaEval Benchmarking Initiative for Multimedia Evaluation. It received participation from 12 teams submitting a total of 54 runs. Highest MAP obtained for the image interestingness data was 0.2336, whereas for video interestingness prediction it was only 0.1815. Although a great deal of approaches were experimented, ranging from standard classifiers and descriptors, to deep learning and use of pre-trained data, the results show the difficulty of this task.

From the experience with this data, we can draw some general conclusions that will help shape future data in this area. Firstly, one should note that generating data and ground truth for such a subjective task is a huge effort and effective methods should be devised to reduce the complexity of annotation. In our approach we took advantage of a pair-wise comparison protocol which was further applied in an adaptive square fashion way to avoid comparing all possible pairs. This has limitation as it still requires a great number of annotators and resulted in a low inter-agreement. A potential improvement may consist on ranking directly series of images/videos. We could also think of crowd-sourcing the key-frames/videos returned by the participants' systems to extract the most interesting samples and evaluating the performances of the systems against these samples only.

Secondly, the source of data is key for a solid evaluation. In our approach we selected movie trailers, due to their Creative Commons licenses which allow redistribution. Other movies are in almost all cases closed content for the community. On the other hand, trailers are edited content which will limit at some point the naturalness of the task, but offer a good compromise given the circumstances. Future improvements could consist of selecting the data as parts of a full movie — a few Creative Commons movies are indeed available. This will require a greater annotation effort but might provide a better separation between interesting and non interesting content.

Thirdly, a clear definition of image/video interestingness is mandatory. The concept of content interestingness is already very subjective and highly user dependent, even compared to other video concepts which are exploited in TRECVID or ImageCLEF benchmarks. A well founded definition will allow for a focused evaluation and disambiguate the information need. In our approach, we define interestingness in the context of selecting video content for illustrating a web site, where interesting means an image/video which would be interesting enough to convince the user to watch the source movie. As a future challenge, we might want to compare the results of interestingness prediction for different use scenarios, or even test the generalization power of the approaches.

Finally, although image and video data was by design specifically correlated, *i.e.,* images were selected as key-frames from videos, results show that actually predicting image interestingness and predicting video interestingness are two completely different tasks. This was more or less proved in the literature, however, in those cases, images and videos were not chosen to be correlated. Therefore, a future perspective might be the separation of the two, while focusing on more representative data for each.

Acknowledgements We would like to thank Yu-Gang Jiang and Baohan Xu from the Fudan University, China, and Hervé Bredin, from LIMSI, France for providing the features that accompany the released data, and Frédéric Lefebvre, Alexey Ozerov and Vincent Demoulin for their valuable inputs to the task definition. We also would like to thank our anonymous annotators for their contribution to building the ground-truth for the datasets. Part of this work was funded under project SPOTTER PN-III-P2-2.1-PED-2016-1065, contract 30PED/2017.

References

1. J. Almeida. UNIFESP at MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
2. J. Almeida, N. J. Leite, and R. S. Torres. Comparison of video sequences with histograms of motion patterns. In *IEEE ICIP International Conference on Image Processing*, pages 3673–3676, 2011.
3. Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.
4. A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, pages 3562–3569. IEEE, 2012.
5. D. E. Berlyne. *Conflict, arousal and curiosity*. Mc-Graw-Hill, 1960.
6. Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1):17–31, 2007.
7. R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: the method of paired comparisons. *Biometrika*, (39 (3-4)):324–345, 1952.
8. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
9. A. Bulling and D. Roggen. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th international conference on Ubiquitous Computing*, pages 455–464. ACM, 2011.
10. C. Chamaret, C.-H. Demarty, V. Demoulin, and G. Marquant. Experiencing the interestingness concept within and between pictures. In *Proceeding of SPIE*, Human Vision and Electronic Imaging, 2016.
11. A. Chen, P. W. Darst, and R. P. Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3):383–400, 2001.
12. S. Chen, Y. Dian, and Q. Jin. RUC at MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
13. S. L. Chu, E. Fedorovskaya, F. Quek, and J. Snyder. The effect of familiarity on perceived interestingness of images. In *Proceedings of SPIE*, volume 8651, pages 86511C–86511C–12, 2013.
14. M. G. Constantin, B. Boteanu, and B. Ionescu. LAPI at MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
15. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, 2005.
16. M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*, 2014.
17. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *IEEE ECCV European Conference on Computer Vision*, pages 288–301. Springer, 2006.

18. 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 *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, 2016.
19. S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.
20. L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008.
21. G. Erdogan, A. Erdem, and E. Erdem. HUCVL at MediaEval 2016: Predicting interesting key frames with deep models. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
22. H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *ACM International Conference on Multimedia*, pages 1017–1026, New York, NY, USA, 2013.
23. M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. van Gool. The interestingness of images. In *ICCV International Conference on Computer Vision*, 2013.
24. M. Gygli, Y. Song, and L. Cao. Video2gif: Automatic generation of animated gifs from video. *CoRR*, abs/1605.04850, 2016.
25. J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.
26. A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.
27. L.-C. Hsieh, W. H. Hsu, and H.-C. Wang. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4309–4313. IEEE, 2014.
28. 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 *ACM International Conference on Multimedia*, 2013.
29. P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.
30. P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, pages 145–152. IEEE, 2011.
31. Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, 177(8):1–13, 2015.
32. Y-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yan. Understanding and predicting interestingness of videos. In *AAAI Conference on Artificial Intelligence*, 2013.
33. T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.
34. S. A. Turner Jr and P. J. Silvia. Must interesting things be pleasant? a test of competing appraisal structures. *Emotion*, 6(4):670, 2006.
35. Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 597–604. IEEE, 2005.
36. Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 419–426. IEEE, 2006.
37. A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, 2015.
38. K. Krippendorff. *Content analysis: An introduction to its methodology, 3rd edition*. SAGE, 2013.
39. V. Lam, T. Do, S. Phan, D.-D. Le, S. Satoh, and D. Duong. NII-UIT at MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.

40. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.
41. J. Li, M. Barkowsky, and P. Le 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*, volume 8648, 2013.
42. L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.
43. C. Liem. TUD-MMC at MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
44. F. Liu, Y. Niu, and M. Gleicher. Using web photos for measuring video frame interestingness. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2058–2063, 2009.
45. Y. Liu, Z. Gu, and Y.-M. Cheung. Supervised manifold learning for media interestingness prediction. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
46. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, (60):91–110, 2004.
47. J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, pages 83–92, New York, NY, USA, 2010.
48. R. R. McCrae. Aesthetic chills as a universal marker of openness to experience. *Motivation and Emotion*, 31(1):5–11, 2007.
49. N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.
50. T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (24(7)):971–987, 2002.
51. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, (42):145–175, 2001.
52. S. Ovadia. Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, 7(5):403–414, 2004.
53. J. Parekh and S. Parekh. The MLPBOON predicting media interestingness system for MediaEval 2016. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
54. S. Rayatdoost and M. Soleymani. Ranking images and videos on visual interestingness by visual sentiment features. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
55. T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber. Coherence progress: a measure of interestingness based on fixed compressors. In *International Conference on Artificial General Intelligence*, pages 21–30. Springer, 2011.
56. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
57. Y. Shen, C.-H. Demarty, and N. Q. K. Duong. Technicolor@MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
58. 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'17, 2017.
59. P. J. Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005.

60. P. J. Silvia, R. A. Henson, and J. L. Templin. Are the sources of interest the same for everyone? using multilevel mixture models to explore individual differences in appraisal structures. *Cognition and Emotion*, 23(7):1389–1406, 2009.
61. M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The mediaeval 2015 affective impact of movies task. In *Proceedings of the MediaEval Workshop*, CEUR Workshop Proceedings, September 2015.
62. M. Soleymani. The quest for visual interest. In *ACM International Conference on Multimedia*, pages 919–922, New York, NY, USA, 2015.
63. M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1):59–76, 2011.
64. B. E. Stein and T. R. Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266, 2008.
65. L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *IEEE ECCV European Conference on Computer Vision*, pages 776–789. Springer, 2010.
66. P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.
67. A. B. Vasudevan, M. Gygli, A. Volokitin, and L. V. Gool. Eth-cvl @ MediaEval 2016: Textual-visual embeddings and video2gif for video interestingness. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
68. J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.
69. B. Xu, Y. Fu, and Y.-G. Jiang. BigVid at MediaEval 2016: Predicting interestingness in images and videos. In *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.
70. Y.-H. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774, 2011.
71. G. N. Yannakakis and J. Hallam. Ranking vs. preference: a comparative study of self-reporting. In *International Conference on Affective Computing and Intelligent Interaction*, pages 437–446. Springer, 2011.