INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Review on Deep Learning for Multi-Modal Video Engagement Prediction

BOGGARAPU SRINIVASULU
Research Scholar
Department of Computer Science and Enigineering
JS UNIVERSITY, Shikohabad ,Uttar Pradesh

Dr. K. VIJAYA BHASKAR
Associate Professor
Department of Computer Science and Enigineering
JS UNIVERSITY, Shikohabad ,Uttar Pradesh

Abstract:

Predicting how engaging media content will be is a crucial but difficult area of study. First, there is the issue that interestingness is both a subjective and a high-level semantic concept, with no universally accepted definition. This adds another layer of complication. This paper details the process of completing the task using state-of-the-art deep learning techniques. We do tests using datasets driven by social factors (Flickr videos) and content factors (videos from the MediaEval 2016 interestingness job). In order to take into consideration the multimodality and temporal aspect of films, we evaluated different architectures of deep neural networks (DNNs), one of which was a novel combination of multiple recurrent neural networks (RNNs), which could process multiple temporal samples concurrently. After that, we looked into various methods for handling imbalanced datasets. Using multimodality, which is the merging of visual and auditory information at a medium level, improved performance on the test. Additionally, we demonstrated that substance interest-ingness is distinct from social interestingness..

Index Terms**—** Flickr videos, social factors, content interestingness, multimodal fusion, DNN.

I INTRODUCTION

Sharing media files like photos and movies is becoming increasingly common in today's fast-paced society. Therefore, in systems like information retrieval and recommendation, the capacity to comprehend such material in order to choose the pertinent ones is crucial. The comprehension of text can be impacted by several notions. Recent studies have focused on higher-level (and possibly less well defined) notions like emotion, popularity, and interestingness[3,4,5,6], in contrast to the extensive prior study on lower-level concepts like visual saliency and aesthetics [1, 2]. This work presents computational models for predicting video interestingness with a singular focus on interestingness. It should be noted that, despite the extensive research on image interestingness, there have only been two publications that have provided benchmark datasets (to the best of our knowledge) [13, 6].

Several fields could benefit from media interest prediction algorithms, including teaching, marketing, content management, and selective encoding. Due to the subjective nature of interest, media sharing websites like Flickr and Pinterest use socially driven metrics like views, tags, comments, user reputations, and viewer profiles to determine the social interest of their content. Consistent with this definition, Liu et al. [11] explored the effect of viewer profiles and suggested using viewer data to determine the attractiveness of images. Assuming that all fashion-related pins on Pinterest are intriguing, Rajani et al. conducted study on predicting the interest of fashion products for online shopping [10]. In [8], Chu et al. looked into how familiarity affected how fascinating people thought pictures were. For a comprehensive overview of previous research on social image interest prediction, see [9]. As for video, Liu et al. [7] assumed that video frames would be considered fascinating if they resembled certain Flickr photographs, therefore they utilised these photos as an indicator to quantify the interestingness of frames in travel videos. In their ground-breaking study on video interestingness, Jiang et al. used the Flickr interestingness API to compile the first video benchmark dataset together with annotations. In this study, they employed support vector machines as a classification method and explored the use of several hand-crafted features.

The use of direct human annotation of content is another area of study. In this method, users are given the freedom to express their personal opinions on the interesting-ness of images [14, 15, 6], image sequences [16], or videos [13, 6]. This is called content interestingness since the annotation is solely dependent on how the material is viewed. The MediaEval 2016 campaign2 just suggested the first benchmark on Predicting Media Interestingness, in keeping with this definition. This work, discussed in depth in [6], has piqued the curiosity of academics and has proven once and for all the importance of comprehending the visual properties of multimedia.

In order to assist users in deciding whether or not to watch a particular piece of content, the task employs two distinct subtasks that are defined in the context of a real-life use case scenario: the presentation of a Video-On-Demand website using movie clips.

Several new contributions to the problem of video interest prediction are detailed in this article. We begin by outlining methods for training computational models that are effective in the face of imbalanced datasets, and then we suggest numerous such models based on modern DNN architectures. We validate the usefulness of multimodal-based systems for the job, which involve combining visual and auditory information at a medium level. Additionally, we provide some new understandings from our research on the differences between socially driven interest and content driven interest, as well as how to anticipate either.

Below is the outline for the remainder of the article. The computational models for video interest prediction and the methods for handling imbalanced data are detailed in Section 2. Results from two datasets' experiments are summed together in Section 3. Finally, in Section 5, we draw conclusions after discussing the suggested systems and their outcomes.

II PROPOSED COMPUTATIONAL MODELS

The workflow of our multimodal interestingness prediction system is comprised of several primary processes, which are as follows: audio modality learning, visual modality learning, multimodal fusion, joint-feature learning, a classification layer with voting for the final predicted label, and finally, the system is trained. The overall workflow can be seen in Figure 1. After you have finished reading this introduction, you will go on to read full explanations of each processing block. The workflows that were taken into consideration for visual-based and audio-based systems are shown in Figure 1. These workflows are compared with monomodal approaches using blue dashed lines and green dashed-dot lines, respectively.

2.1. Low-level visual and audio features

By taking it from AlexNet's CaffeNet model 3, we were able to incorporate a well-known CNN feature into our visual design [17]. This CNN model that has been pre-trained selects the coefficients of the last dense layer (fc7) before to the softmax in order to obtain its 4096-size feature. When the video frames are split along the centre, they are compressed until they are 227 pixels in size. This is done in order to make the video frames smaller enough to fit inside the dimensions of 227 × 227. Furthermore, in order to guarantee that the training plan is followed, their means are subtracted as part of the input normalisation process.

A characteristic of each sound frame is comprised of sixty Mel-frequency cepstral coefficients (MFCC)[18] as well as the first and second derivatives of these coefficients across a window of eighty milliseconds. There are 180 bytes in length for the audio feature that was built. For the purpose of normalising the input, the means of all of the MFCC feature vectors are located and then subtracted from one another. The number of feature vectors for the audio signal and the visual representation are same. This is due to the fact that the overlapping windows for the short-term Fourier transform of the audio signal are centred around each frame of the video. It is because of this that the visual and auditory representations continue to be in sync with one another.

2.2. Learning of features that are temporally driven and synthesis of several modalities

We use the long-short term memory (LSTM), which is a well-known architecture for simulating long-term associations [19], for the two stages of advanced monomodal feature learning. This is due to the fact that temporal development is a substantial component of a visual signal. During the training process, it is anticipated that LSTM layers would be able to identify changes that have occurred over a period of time, as represented in low-level frame-based feature vectors. These feature vectors include those used by CNN and

MFCC. Not only that, but we also train extremely deep CNNs by using the newly proposed residual network (ResNet) [20], which is what led us to employ this LSTM layer architecture for learning the residual functions from the inputs of the LSTM layers. It was not possible for us to locate any prior study that dealt with movies and used ResNet blocks in conjunction with LSTM layers. In addition, we include a multilayer perceptron (MLP) layer at the beginning of the visual block in order to achieve a balance between the two different sources of information. The sizes of the graphic components are decreased by this layer so that they are more closely aligned with the proportions of the audio characteristics. While the installation was being carried out, we made a number of modifications to the features of both the visual and aural modes. There are many different sorts of parameters, some examples of which are activation functions, dropout, input/output sizes, and layer types.

Following the completion of this higher-level modelling of a single modality, the multimodal fusion stage involves the fusing of both modes in order to produce an image that contains multiple modes.

2.3. Multimodal feature learning

Our objective is to merge the higher-level representations that come from the two different modalities that were taught independently of one another. The sections 2.3.1 and 2.3.2 will provide a comprehensive overview of our inquiry into two different DNN architectures that we use for this particular purpose.

2.3.1. (LSTM/Resnet)-based architecture

LSTM is once again the technique that is considered to be the most effective approach for dealing with the temporal connection of the multi-modal feature vectors. It is our intention to establish a connection between LSTM and the ResNet architecture for the single-modal branches that were covered in Section 2.2. This will prevent the training process from overfitting the data. During the course of execution, there was a significant amount of reconsidering of choices about network design and parameters.

2.3.2. Proposed n-RNN-based architecture

We provide a unique design that improves temporal modelling by using multiple recurrent neural network (RNN) nodes. This design is in addition to the designs that have been considered to be state-of-the-art in the past. The W, U, and V weights that were determined during training are used by each and every one of these n RNN nodes. This design accepts n input samples xi,t (where i = 1,..., n) and utilises n internal states si,t to generate n internal outputs yi,t for each time event t. The n internal outputs are also referred to as yi,t.:

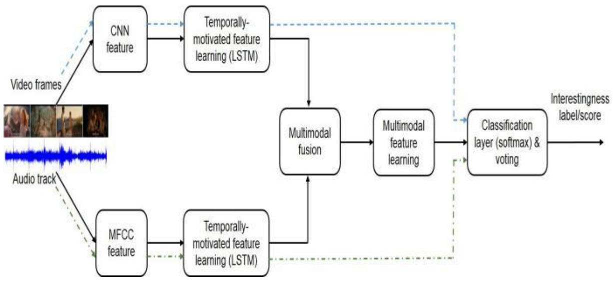

Video frames
Audio track
CNN feature
MFCC feature
Temporally-motivated feature learning (LSTM)
Temporally-motivated feature learning (LSTM)
Multimodal fusion
Multimodal feature learning
Classification layer (softmax) & voting
Interestingness label/score

Fig. 1 Description Prediction of video interest using suggested computer models. Here we can see the process for our multimodal method as black arrows, and for visual-based and audio-based systems, respectively, as blue dashed lines and green dash-dot lines, representing monomodal workflows.

In other words, the nonlinear activation functions of f and g are being discussed here. For the purpose of obtaining the ultimate temporal output yнт [21] from the internal outputs yi and t, a time-delayed neural network (TDNN) is often used. As shown in Figure 2, the standard RNN unfolding as well as the unfolding of the proposed design are both shown. It is possible that the suggested design will have an easier difficulty comprehending the idea of a single instant in time if it continues to study a large number of samples concurrently. It is possible for it to get more comprehensive so long as it does not have to recall each and every information. In addition, we are of the opinion that the training process is sufficiently clever to ascertain the most effective method for combining the internal discoveries yi,t. This is the reason why we chose to use TDNN rather than a simple pooling methodology. There are a number of ways in which TDNN is superior than rival DNN systems such as MLP. One of these ways is that it enables faster learning.s.

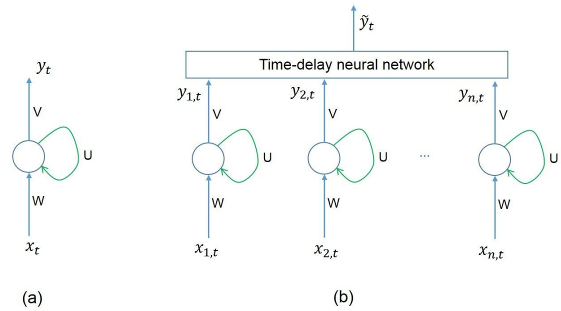


Fig. 2: Comparative analysis of a conventional RNN (a) and (b) our recently developed design, which incorporates several RNN nodes.

2.3.3 Classification layer and voting

The results of the frame-based interestingness prediction are formed by feeding the outputs of the multimodal feature learning block into a logistic regression layer, namely softmax [22]. This layer is responsible for generating the findings. For the purpose of determining the ultimate outcome for the video's interestingness prediction, we use a voting technique that involves averaging these results.

2.4. Practicing with datasets that are not balanced

We looked at two different methods in order to deal with the fact that one of our datasets was quite tiny and lacked balance.:

Up-sampling: Throughout the training process, each and every sample from the minority class is repeated. When it comes to our situation, we make use of engaging movies on various occasions during each training period.

Random sampling: Videos that are intriguing and videos that are not interesting are sorted into two groups. The samples are then selected at random from both sets with a predetermined probability while the training is being conducted.

IV EXPERIMENTS

4.1 Datasets

In a sequential fashion, two datasets were used to train the aforementioned learning-based systems. In [13], the first dataset was proposed; from here on out, we'll refer to it as Jiang's dataset. It uses the Flickr interestingness API and consists of 1200 films, each with a total running time of 20 hours and an average duration of one minute. Keep in mind that this API relies on user-generated content interactions, therefore the social interest annotation it generates might not always match up with a true content interest groundtruth. The videos were culled by conducting a search using a set of fifteen keywords. The top ten percent of the results were deemed intriguing, while the bottom ten percent were considered uninteresting. Keep in mind that Jiang's dataset is well-rounded and includes a variety of content kinds, not all of which are professional. Because of constraints in the hardware, we had to resort to using 60-frame random samples while training with this dataset. We used 70% for training, 15% for validation, and 15% for testing on this dataset.

The Mediaeval 2016 video dataset is the second dataset [6]. On average, each of the 5,054 shots used for development and 2,342 shots used for testing lasts 1 second. In order to extract these scenes, 78 movie trailers that resemble Hollywood were manually segmented, indicating that the content was professional. With only 8.3% of the development set and 9.6% of the test set containing interesting content, this second dataset is severely imbalanced. This time, content-driven interestingness assessment resulted from annotations that were based entirely on the content. We divided the development set in half, allocating 80% for training and 20% for validation, in order to optimise the computational models.

4.2 Systems and prediction results

In Section 3.2.1, we detail our work using Jiang's dataset to forecast interest based on social factors, and in Section 3.2.2, we detail our work using the MediaEval dataset to forecast interest based on content factors.

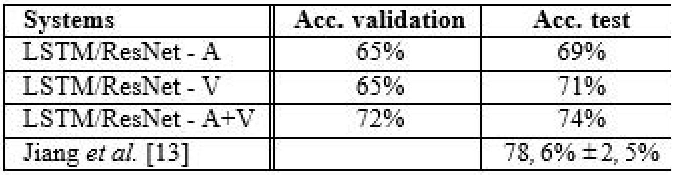
4.2.1 Observations made with Jiang's dataset

The results of our testing with a variety of parameter choices for the multimodal and monomodal approaches are shown in Figure 1. The activation function, the number of DNN layers, the kind of layer, the size of the layer, and a great deal of other factors were included in this set of variables.

A fundamental architecture that consisted of one LSTM layer, ReLu activation, and dropout=0.5 in the temporally-motivated feature learning block—prior to the softmax classification layer and the voting section—and had an output size of 180 performed well when the audio modality was used on its own.

The process of selecting the most suitable layout for the video mode was somewhat more challenging. Among the components of the model was an MLP that decreased the size of the input features from 4096 to 1024, an LSTM layer that had an output size of 256, and an extra 256 LSTM layer that was included inside a ResNet arrangement. Every single one of them had ReLu activation and dropout levels that were equal to 0.5. We were able to get the same results even with a simplified architecture that consisted of just two layers: an MLP with output dimensions of 1024 by 0.5 dimensions and an LSTM with output dimensions of 256 by 0.5 dimensions.

From what can be seen in Figure 1, the multimodal processing component was the one that received the most attention. We made sure to keep the design of the two monomodal branches, which are the fundamental components of time-driven learning features, as basic as possible inside the multimodal framework. We chose the second-best and most straightforward method for video processing since it only required one LSTM layer (with an output dim of 256) and one MLP layer (with an output dim of 1024). A design that just has one LSTM layer is the one that we recommend the most for audio. The best configuration for the multimodal feature learning block was equally straightforward: it was triggered by ReLu, and it consisted of two layers of long short-term memory (LSTM) with output sizes of 436 and 218, respectively. All of the information that was collected from the multimodal and single-modal systems is shown in Table 1. The performance of the video modality on the test set is superior than that of the audio modality by a margin of 2%. Following that, the multimodal feature learning block makes use of the multimodal LSTM/ResNet technique in order to ascertain the most accurate values for both the test and validation sets. The assertion that multimodality leads to increased productivity in the workplace is given more weight as a result of this. Due to the fact that we did not train and test on the whole video samples, as Jiang et al. did in [13], it would be inappropriate to compare our work to what they performed.



Systems	Acc. validation	Acc. test
LSTM/ResNet - A	65%	69%
LSTM/ResNet - V	65%	71%
LSTM/ResNet - A+V	72%	74%
Jiang *et al.* [13]		78, 6% ±2, 5%

Table 1: The results from Jiang's dataset show how well the prediction of how interesting a movie is worked (in percentages). V: video; A: sound

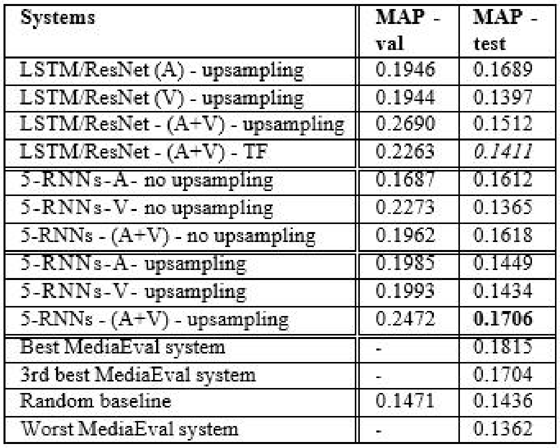
4.2.2 How the MediaEval dataset performed

In the case of Jiang's dataset, the fundamental one-modal designs were used. Our investigation of the MediaEval dataset focused on the design of the multimodal feature learning block as its primary location of concentration. It was noticed by us that modifying the number of LSTM/ResNet layers did not have any impact on the performance of either the complicated or the simple techniques. There is a good chance that the dataset was too small to cause this. After much deliberation, we decided to use a single ResNet structure for this block. This structure would consist of one LSTM layer, 436 output pixels, and a dropout value of 0.5.

Additionally, we put our innovative approach to conveying time via the use of both multi-modal and single-modal settings to the test. Due to the fact that the RNNs were set up in a certain order, our RNN-based system was able to concurrently analyse five consecutive time samples. In the process of monomodal learning, we simply replaced the layers of LSTM and potential ResNet that had been evaluated in the past with our innovative temporal modelling. The LSTM/Resnet design was replaced with this 5-RNN-based architecture; however, the single-modal branching of the multimodal feature learning block was maintained.

According to the information presented in Section 2.5, it has been shown that the use of data augmentation methods, such as resampling or upsampling the raw data, may improve performance and correct the imbalance that exists within the MediaEval dataset. To such an extent that we tested with a wide variety of layouts, some of which included sampling and upsampling, while others did not, and we also experimented with different values for the sampling levels and additional sample. Increasing the pace at which you resampled or upsampled was always the best course of action. In light of the facts that we obtained, we decided to keep the upsampling factor at 9 for all of the designs. There are nine instances in which the training process takes use of compelling examples.

The findings of each system are shown in Table 2, which compares them to the official performance measure for the MediaEval program, which is the mean average accuracy (MAP) scores. In spite of the existence of the recently introduced n-RNN-based architecture, multiple samples were employed in order to guarantee that the training samples were dispersed in an even manner. According to the findings, multimodal techniques are superior to their monomodal counterparts in their performance the vast majority of the time. When it comes to learning features, this is true regardless of whether the multimodal approach use LSTM/ResNet or our innovative n-RNN-based architecture. When compared to Jiang's dataset, this job is far more difficult to complete on the MediaE-val dataset. This is before we even take into consideration the range of evaluation metrics that are available. The straightforward explanation for this is that the accuracy values are greater than the MAP values. Furthermore, we proved that our new model, which is based on n-RNN, beat the most recent LSTM/ResNet architecture on both the test and validation sets. This showed that our model is superior to the existing design. We also present the results of numerous official entries to the MediaEval Predicting Conversation Interestingness challenge so that you may get a sense of how well they do in contrast to the current state of the art. They are displayed here for your convenience. Particularly noteworthy is the fact that our revised method surpassed both the randomly assigned baseline, in which samples are split into two groups, and the third-best approved procedure, which was chosen from a total of twenty-eight entries.



Systems	MAP - val	MAP - test
LSTM/ResNet (A) - upsampling	0.1946	0.1689
LSTM/ResNet (V) - upsampling	0.1944	0.1397
LSTM/ResNet - (A+V) - upsampling	0.2690	0.1512
LSTM/ResNet - (A+V) - TF	0.2263	*0.1411*
5-RNNs-A- no upsampling	0.1687	0.1612
5-RNNs-V- no upsampling	0.2273	0.1365
5-RNNs - (A+V) - no upsampling	0.1962	0.1618
5-RNNs-A- upsampling	0.1985	0.1449
5-RNNs-V- upsampling	0.1993	0.1434
5-RNNs - (A+V) - upsampling	0.2472	**0.1706**
Best MediaEval system	-	0.1815
3rd best MediaEval system	-	0.1704
Random baseline	0.1471	0.1436
Worst MediaEval system	-	0.1362

Table 2 If MAP is 1, samples that were expected are ranked first, and if MAP is 0, all samples that were not expected are ranked first. These are the prediction results for the MediaEval validation and test datasets. We increased the size of a factor 9 when we said that (training, validation) = (80%, 20%). A: sound; V: video; TF: learning from Jiang's model transfer.

V DISCUSSION

During the course of our testing, we were able to arrive at a few findings. In light of the findings shown in Tables 1 and 2, it is important to point out that multimodal systems performed better than monomodal ones. The only two situations in which multimodal systems did worse were when we used LSTM-based systems on the MediaEval test set and when we used our new n-RNN-based models without upsampling. It is not possible to consider the multimodal system in the first case to be very generalisable because of the small amount of the dataset. The fact that the dataset is very huge and does not include any upsampling, which results in a mismatch, is still another argument that may be given for the second case.

We discovered things that were completely different between the two samples. This might be true in a variety of different ways across the board. Due to the limited size of the MediaEval dataset or the difficulties in distinguishing between content-driven interest and social-driven interest, our results may not be generalisable. This is one of the potential outcomes that may occur. As a result of the conversation, it is abundantly clear that the two ideas are quite distinct from one another. Detailed information may be found in Table 2. The results of our tests on transfer learning supported this assertion. We used Jiang's dataset to train a deep neural network (DNN)-based multimodal model, and then we used that model to the MediaEval dataset in order to estimate consumer preferences. It would seem that there is a cognitive gap, as shown by the unsatisfactory results (MAP = 0.1411). Two different sorts of information are presented here. In the first, you will discover material that was made by average individuals, and in the second, you will find information that was developed by trained professionals. There is a possibility that this is a contributing element. It is likely that the overall poor performance might be attributed to the large number of blurry, small photos that are included in the MediaEval collection as well as the comments that are linked with them. Due to the fact that interest is a matter of personal preference, the opinions of users may not always give an accurate indication of the quality of content-based films. This gives rise to the need for further queries.

Our innovative n-RNN-based structure obtained a best MAP value that was much higher than the baseline, which indicates that the system did comprehend the interestingness principle. This is despite the fact that the sample counts may be insufficient to train sophisticated DNN structures. In comparison to state-of-the-art deep neural network designs that are based on LSTM and Resnet, our cutting-edge n-RNN-based structure is superior in terms of its ability to concurrently investigate several temporal data sets. The findings of this study suggest that our technique to data temporal modelling is applicable to a wider range of situations.

VI CONCLUSION

Within the scope of this study, we provide a general computational model that is capable of predicting the level of engagement that will be shown by video content by using cutting-edge deep learning architectures. We evaluate its effectiveness by applying it to two different video datasets: one uses annotations based on the content's interestingness, while the other uses annotations based on social factors. Our findings demonstrate that multimodal-based approaches that include mid-level audio and visual feature fusion perform much better than monomodal-based systems when applied to both datasets. Based on our findings, we have come to the realisation that there is a significant distinction between the social interestingness and the content interestingness. We hope that by collecting a larger dataset with reliable annotation, which is our objective for the next research, we will be able to get a deeper understanding of the inherent appeal of a movie. In order to take into consideration the subjective aspect of the idea even further, we will furthermore concentrate our study on simulating the contextual interestingness of scenarios.

REFERENCES

[1] Simone Frintrop, Erich Rome, and Henrik I. Christensen, “Computational visual attention systems and their cognitive foundations: A survey,” *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 1–39, Jan. 2010.

[2] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah, “A framework for photo-quality assess- ment and enhancement based on visual aesthetics,” in *Proc. of ACM International Conference on Multimedia(MM), Florence, IT*, 2010, pp. 271–280.

[3] U. Rimmele, L. Davachi, R. Petrov, S. Dougal, and E. A. Phelps, “Emotion enhances the subjective feeling of re- membering, despite lower accuracy for contextual de- tails,” *Psychology Association*, 2011.

[4] L. Mai and G. Schoeller, “Emotions, attitudes and mem- orability associated with tv commercials,” *Journal of Targeting, Measurement and Analysis for Marketing*, pp. 55–63, 2009.

[5] A. Khosla, A. Sarma, and R. Hamid, “What makes an image popular?,” in *Proc. International conference on World Wide Web*, 2013, pp. 867–876.

[6] Claire-Hélène Demarty, Mats Sjberg, Bogdan Ionescu, Than-Toan Do, Hanli Wang, Ngoc Q. K. Duong, and Frédérique Lefebvre, “Mediaeval 2016 predicting media interestingness task,” in *Proc. MediaEval 2016 Workshop, Netherlands*, 2016.

[7] Feng Liu, Yuzhen Niu, and Michael Gleicher, “Using web photos for measuring video frame interestingness,” in *Proceedings of the 21st International Jont Confer- ence on Artifical Intelligence*, San Francisco, CA, USA,2009, IJCAI’09, pp. 2058–2063.

[8] Sharon Lynn Chu, Elena Fedorovskaya, Francis Quek, and Jeffrey Snyder, “The effect of familiarity on perceived interestingness of images,” 2013, vol. 8651, pp. 86511C–86511C–12.

[9] Xesca Amengual, Anna Bosch, and Josep Llu´ıs de la Rosa, *Review of Methods to Predict Social Image Interestingness and Memorability*, pp. 64–76, Springer, 2015.

[10] N. Rajani, K. Rohanimanesh, and E. Oliveira, “Identifying interestingness in fashion e-commerce using pinterest data,” 2015.

[11] B. Liu, M. P. Kato, and K. Tanaka, “Estimating interestingness of iimage based on viewer data,” in *DEIM Forum*, 2015.

[12] 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 Intelli- gence*, 2013.

[13] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Gool, “The interestingness of images,” in *ICCV International Conference on Computer Vision*, 2013.

[14] Michael Gygli and Mohammad Soleymani, “Analyzing and predicting gif interestingness,” in *Proceedings of the 2016 ACM on Multimedia Conference*, New York, NY, USA, 2016, MM ’16, pp. 122–126, ACM.

[15] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool, “Visual interestingness in image sequences,” in *Proceedings of the 21st ACM*

International Confer- ence on Multimedia, New York, NY, USA, 2013, pp. 1017–1026.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hin- ton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges,
L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[17] Steven Davis and Paul Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[18] Sepp Hochreiter and Jürgen Schmidhuber, “Long short- term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in*arXiv prepring arXiv:1506.01497*, 2015.

[20] Geoffrey E. Hinton Kevin J. Lang, Alex H. Waibel, “A time-delay neural network architecture for isolated wordrecognition,” *Neural Networks*, vol. 3, pp. 23–43, 1990.

[21] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.