Deep learning for multimodal-based video interestingness prediction

Yuesong Shen, Claire-Hélène Demarty, Ngoc Q. K. Duong

DEEP LEARNING FOR MULTIMODAL-BASED VIDEO INTERESTINGNESS PREDICTION

*Yuesong Shen**

Technische Universität München
Munich, Germany
yuesong.shen@tum.de

Claire-Hélène Demarty, Ngoc Q.K. Duong

Technicolor, Rennes, France
claire-helene.demarty@technicolor.com
quang-khanh-ngoc.duong@technicolor.com

ABSTRACT

Predicting interestingness of media content remains an important, but challenging research subject. The difficulty comes first from the fact that, besides being a high-level semantic concept, interestingness is highly subjective and its global definition has not been agreed yet. This paper presents the use of up-to-date deep learning techniques for solving the task. We perform experiments with both *social-driven* (*i.e.,* Flickr videos) and *content-driven* (*i.e.,* videos from the MediaEval 2016 interestingness task) datasets. To account for the temporal aspect and multimodality of videos, we tested various deep neural network (DNN) architectures, including a new combination of several recurrent neural networks (RNNs), to handle several temporal samples at the same time. We then investigated different strategies for dealing with unbalanced datasets. Multimodality, as the mid-level fusion of audio and visual information, brought benefit to the task. We also established that social interestingness differs from content interestingness.

Index Terms— Video interestingness prediction, social interestingness, content interestingness, multimodal fusion, deep neural network (DNN), MediaEval 2016.

1. INTRODUCTION

In our fast moving world, the amount of shared data such as images and videos is growing exponentially. Thus the ability to understand such content so as to select the relevant ones plays a key role in *e.g.,* information retrieval and recommendation systems. Different concepts may intervene in the understanding of content. While the lower level concepts, such as visual saliency and aesthetics, have been studied for a long time [1, 2], some recent research targets higher level (and potentially less well-defined) concepts such as emotion, popularity and interestingness [3, 4, 5, 6]. Focusing solely on interestingness, this paper proposes computational models for video interestingness prediction. Note that while image interestingness has been widely studied in the literature [7, 8, 9, 10, 11, 12], video interestingness has been much less investigated with, to our knowledge, only two publications providing benchmark datasets [13, 6].

Predicting media interestingness has several potential applications in, *e.g.,* education, advertising, selective encoding or content management. As interestingness is highly subjective, media sharing websites such as Flickr and Pinterest extract interestingness of their content, to which we refer as *social interestingness* in the following, based on social-driven criteria such as the number of views, tags, comments, user reputations and viewer's profiles[1]. In line with this definition, Liu *et al.* [11] proposed to estimate the interestingness of images based on viewer data and investigated the impact of viewer's profiles. Rajani *et al.* performed research in predicting interestingness of fashion products for online shopping, where they assumed that all pins related to fashion on Pinterest are interesting [10]. Chu *et al.* investigated the effect of familiarity in the perceived interestingness of images in [8]. Readers are referred to [9] for a review of existing work in predicting social image interestingness. Moving toward video, Liu *et al.* [7] used web photos such as Flickr images as an indicator to measure the interestingness of frames in travel videos, with the assumption that video frames are interesting if they are found similar to some Flickr photos. Jiang *et al.* presented a pioneer work on video interestingness where a first video benchmark dataset and its annotation are also collected from Flickr interestingness API. In this work, they investigated the use of different hand-crafted features, and used support vector machines as classification technique.

Another axis of research uses direct human annotation of content where users are asked to freely judge the interestingness of images [14, 15, 6], image sequences [16], or videos [13, 6] based on their own opinion. We refer to this as *content interestingness* since the annotation is purely based on the perceived content itself. In line with this definition, a first benchmark on Predicting Media Interestingness was recently proposed in the MediaEval 2016 campaign[2]. The task which is described in details in [6], has raised a huge interest in the research community, confirming the need for understanding the perceptual characteristics of multimedia content.

*The author performed the work while working at Technicolor.

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

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

978-1-5090-6067-2/17/$31.00

The task targets content interestingness of both images and videos, thanks to two separated subtasks, defined in the context of a real use case scenario: the illustration of a Video-On-Demand web site by movie excerpts, so that it helps a user decide whether he/she is interested in watching a given content.

This paper states several contributions to the emerging task of predicting video interestingness. First we propose several efficient computational models based on up-to-date DNN architectures, and describe techniques for dealing with unbalanced datasets while training such models. We confirm the benefit of multimodal-based systems, as the mid-level fusion of audio and visual features, to the task. We also bring some insights in the study of the two concepts: social driven interestingness and content driven interestingness, and on how their prediction differs.

The rest of the paper is organized as follows. In Section 2 we present the proposed computational models for video interestingness prediction as well as the considered techniques used for dealing with unbalanced data. Experiment results on two datasets are summarized in Section 3. We discuss the proposed systems and their results in Section 4, and finally conclude in Section 5.

2. PROPOSED COMPUTATIONAL MODELS

The general workflow for our multimodal interestingness prediction system (see Fig. 1) consists of several major steps: audio modality learning, visual modality learning, multimodal fusion and joint-feature learning, followed by a classification layer with voting for the final predicted label. Each processing block is described in details in the following subsections. Note that for comparison with monomodal approaches, the considered workflows for visual-based and audio-based systems are presented with blue dash lines and green dash-dot lines, respectively in Fig. 1.

2.1. Low-level visual and audio features

For the visual modality, we use the well-known convolution neural network (CNN) feature extracted from the CaffeNet model [3], a variant of AlexNet [17]. In this pre-trained CNN model, the coefficients of the last dense layer (fc7) before the softmax are chosen, yielding a feature of size 4096. In order to fit the input size of CaffeNet (*i.e.,* 227×227), the video frames are first re-scaled so that their smaller dimension is 227, then center-cropped. Additionally, their mean values are subtracted as an input normalization to meet the protocol upon which the model was trained.

For the audio modality, we extract a vector of 60 Mel-frequency cepstral coefficients (MFCC) [18], together with its first and second derivatives over a window size of 80ms, as the feature for each audio frame. The resulting audio feature size is therefore 180. The mean values over the whole video are calculated and subtracted from each MFCC feature vector in order to perform input normalization. Note that, to synchronize audio and visual representations, the overlapping windows for the short-term Fourier transform of the audio signal are centered around each frame of the video, resulting in the same number of audio and visual feature vectors.

[3] https://github.com/BVLC/caffe/wiki/Model-Zoo#caffenet-fine-tuned-for-oxford-flowers-dataset

2.2. Temporally-motivated feature learning and multimodal fusion

As temporal evolution is an important property of a video signal, for the two steps of higher-level monomodal feature learning, we exploit the long-short term memory (LSTM) [19] architecture, which is well-known for the modeling of long-term dependencies. We expect that LSTM layers will learn a representation that accounts for the temporal evolution encoded in the low-level frame-based feature vectors (*i.e.,* CNN and MFCC). Besides, motivated by the effectiveness of the recently proposed residual network (ResNet) for training very deep CNNs [20], we exploit this architecture for LSTM layers so as to learn the residual functions with reference to the LSTM layer inputs. To our knowledge, this is the first work to use the ResNet blocks with LSTM layers for a video. Besides, we incorporate a multilayer perceptron (MLP) layer at the beginning of the visual block to reduce the visual feature size to the same order as the audio feature size for balancing the two modalities. Note that in the implementation, we have used different parameter settings (layer types, layer input/output sizes, dropout, activation functions, *etc.,*) for both the visual and audio modalities.

After this monomodal higher-level feature modeling, both modalities are simply concatened, in the multimodal fusion step, to form a multimodal feature representation.

2.3. Multimodal feature learning

From the separated learning of modal-specific features, we expect to jointly build a higher-level representation from the two modalities. We investigate two DNN architectures for this purpose as it will be described in Sections 2.3.1 and 2.3.2.

2.3.1. (LSTM/Resnet)-based architecture

To account for the temporal correlation of the derived multimodal feature vectors, LSTM is again a natural choice at this step. Like what is proposed for the monomodal branches in Section 2.2, we couple LSTM with the ResNet architecture so as to potentially avoid overfitting during training. Again in the implementation, several parameter settings and network architectures have been explored.

2.3.2. Proposed n-RNN-based architecture

In addition to the previous up-to-date architectures, we propose another architecture exploiting several recurrent neural network (RNN) nodes altogether in order to attempt for a better temporal modeling. These n RNN nodes share the same weights W, U, V, which are learned during training. At each time instance t, this architecture takes into account n input

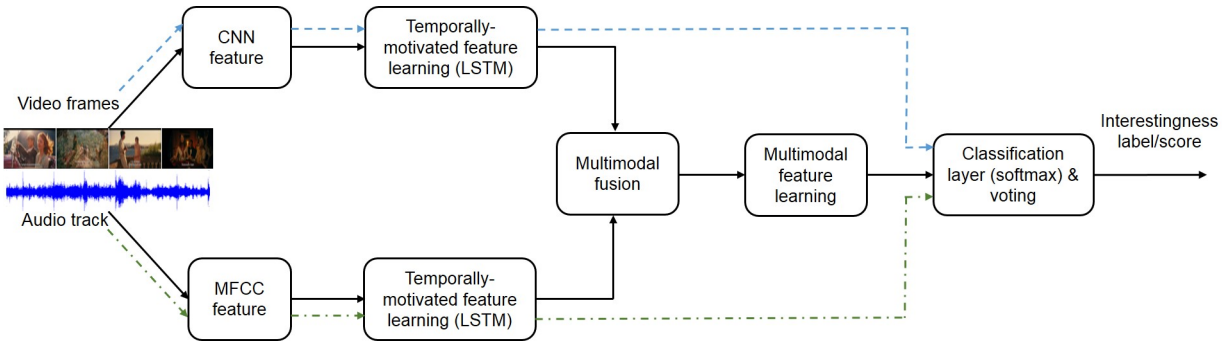

Fig. 1: Proposed computational models for video interestingness prediction. Black arrows represent the workflow for our multimodal approach, whereas blue dash lines and green dash-dot lines represent monomodal workflows for visual-based and audio-based systems, respectively.

samples $x_{i,t}$ $(i = 1, ..., n)$ together with n internal states $s_{i,t}$ to produce the corresponding n internal outputs $y_{i,t}$ as in the following:

$$x_{i,t} = x_{t-n+i} \quad (1)$$

$$s_{i,t} = f(Wx_{i,t} + Us_{i,t-1}) \quad (2)$$

$$y_{i,t} = g(Vs_{i,t}), \quad (3)$$

where f, g are the non-linear activation functions. Then the internal outputs $y_{i,t}$ are fed into a time-delayed neural network (TDNN) [21] to produce the final temporal output $\tilde{y}_t$. The unfolding of the proposed architecture is compared to the traditional RNN in Fig. 2. By taking into account multiple successive samples simultaneously, the proposed architecture would potentially better grasp the idea of time *moment* in the video where the learned output $\tilde{y}_t$ could represent a higher level of abstraction instead of trying to memorize every detail. Note also that we use TDNN instead of a simple pooling strategy, with the hope that the best combination of the internal outputs $y_{i,t}$ can be smartly learned during the training. Also the training of TDNN is actually much faster as compared to other DNN architectures such as MLP.

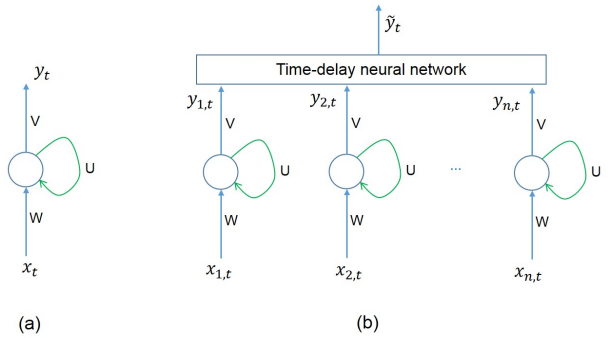

Fig. 2: Comparison between a standard RNN (a) and our new architecture with several RNN nodes (b).

2.4. Classification layer and voting

Outputs from the multimodal feature learning block are fed into a logistic regression layer (*i.e.,* softmax [22]) to produce frame-based interestingness prediction results. Finally these results are averaged, as a voting strategy, to obtained the final interestingness prediction result for the whole video.

2.5. Training with unbalanced datasets

To cope with the small size and unbalance of one of our datasets, we investigated two techniques:

Up-sampling: Every sample from the minority class is repeated during training. In our case, interesting videos are used multiple times during each training epoch.

Random sampling: Interesting and non-interesting videos are separated in two sets. During training, samples are then randomly picked up from both sets with a given probability.

3. EXPERIMENTS

3.1. Datasets

The above learning-based systems were trained successively with two datasets. The first dataset, which we call Jiang's dataset in the following, was proposed in [13]. It is composed of 1200 videos, with an average duration of one minute and 20 hours in total, from Flickr interestingness API. It must be noticed that this API is based on social interactions around the content, meaning that there may be some discrepancy between the resulting social interestingness annotation and a genuine content interestingness groundtruth. Those videos were collected by searching with a set of 15 keywords and keeping the top 10% videos as the interesting subset and the bottom 10% as non interesting samples. Note that Jiang's dataset is equally balanced and contains diverse types of content, non necessarily professional. During our training with this dataset, random video samples of 60 frames were used, due to hardware capability limitations. We split this dataset into 70%, 15%, 15% for training, validation, and testing, respectively.

The second dataset is the Mediaeval 2016 video dataset [6]. It contains 5,054 shots for the development data, and 2,342 shots for the test data, with an average duration of 1s per shot. These shots were extracted thanks to a manual segmentation of 78 Hollywood-like movie trailers, hence from professional content. This second dataset is highly unbalanced with 8.3% (resp. 9.6%) of interesting content for the development set (resp. test set). This time, the annotation was solely based on the content itself, therefore leading to content-driven interestingness assessment. For optimizing the computational models, we split the development set into 80% and 20% for training and validation, respectively.

3.2. Systems and prediction results

We present our study for predicting social-driven interestingness with Jiang's dataset in Section 3.2.1, and content-driven interestingness with the MediaEval dataset in Section 3.2.2.

3.2.1. Results with Jiang's dataset

We experimented different parameter settings (number of DNN layers, layer type and size, activation function, *etc.,*) to chose the most performing architecture for each monomodal and multimodal approaches as proposed in Fig. 1.

For the audio modality alone, a simple architecture of one single LSTM layer of output size 180 (with ReLu activation, dropout=0.5) in the temporally-motivated feature learning block before the classification layer (softmax) and voting part happened to perform well.

For the video modality alone, a slightly more complicated architecture was selected as the most performing one, with one MLP which reduces the size of the input features from 4096 to 1024, one LSTM layer of output size 256 and one ResNet structure with another LSTM layer of output size 256, again all with ReLu activation and dropout=0.5. It should be noted that another simple architecture with only one MLP (output dim=1024, dropout=0.5), followed by one single LSTM layer (output dim=256, dropout=0.5) gave comparable results.

For multimodality, we focused our testing of different architectures on the multimodal processing part only as proposed in Fig. 1, while keeping simple architectures for the two monomodal branches (*i.e.,* temporally-motivated feature learning blocks). For the audio, we elected the above most performing architecture with one single LSTM layer, while for the video branch, we chose the second best architecture, also for simplicity, with one MLP layer (output dim=1024) followed by only one LSTM layer (output dim=256). For the multimodal feature learning block, the best performing archictecture was also quite simple: only two LSTM layers of output sizes 436 and 218, respectively (with ReLu activation). Table 1 shows the results obtained with monomodal and multimodal systems. It can be seen that on the test set, the video modality works slightly better than the audio modality with an increase of 2% of accuracy. The multimodal LSTM/ResNet approach for the multimodal feature learning block results in turn in the highest accuracy values for both the validation and test sets. This confirms the expectation that multimodality brings benefit to the task. Note that the comparison with the work from Jiang *et al.* is not completely fair as we did not perform the training and testing on the complete video samples as in [13].

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]		$78,6\% \pm 2,5\%$

Table 1: Results on Jiang's dataset in terms of accuracy ($\%$) for the prediction of video interestingness. A: audio; V: video

3.2.2. Result with the MediaEval dataset

For the MediaEval dataset, we focused our investigation on the architecture of the multimodal feature learning block, while keeping the simple monomodal architectures from Jiang's dataset. By varying the number of LSTM/ResNet layers, we found again that simple architectures were in general performing as good as the more complex ones, probably due to the small size of the dataset. In the end, we selected a single ResNet structure with one LSTM layer of output size 436 for this block (dropout=0.5).

We also tested our new structure for temporal modeling for both the monomodal and multimodal cases. Each time our RNN-based structure contained 5 RNNs successively arranged so as to process 5 successive temporal samples at the same time. Compared to the previously tested monomodal systems, we simply replaced the LSTM and potential ResNet layers by this new temporal modeling in the monomodal learning step. For the multimodal approach, we kept the monomodal branches as they were before and replaced the LSTM/Resnet architecture in the multimodal feature learning block by this 5 RNN-based architecture.

To cope with the unbalance of the MediaEval dataset, data augmentation thanks to either resampling or upsampling strategies of the input data, as explained in section 2.5, has proven to bring benefits to the performances. Indeed, we tested the different architectures with and without resampling and upsampling techniques (and with different resampling probabilities and upsampling factors). Each time, performances were increased thanks to resampling or upsampling, the best performance being achieved with an upsampling of factor 9 (meaning that interesting samples are repeated 9 times during training), which we decided to keep for all architectures.

Table 2 presents the results we obtained with the different systems in terms of mean average precision (MAP) values, the official evaluation measure in the MediaEval campaign. As can be seen, upsampling to balance the training samples is helpful also for the new n-RNN-based architecture. We

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: Prediction results on the MediaEval validation and test datasets in terms of MAP values ($MAP = 1$: expected samples are ranked first; $MAP = 0$: all non expected samples are ranked first). Upsampling of a factor 9 when mentionned, (training, validation) = (80%, 20%). A: audio; V: Video. TF: transfer learning from Jiang's model.

also observe that, once again, multimodal approaches (either with LSTM/ResNet or with our new n-RNN-based architecture for the multimodal feature learning) perform generally better than monomodal ones. These results also show that the task happens to be much more difficult on the MediaEval dataset than on Jiang's dataset, not taking into account the change of evaluation metrics, as MAP values are somewhat lower than accuracy values. Furthermore, we showed that our new n-RNN-based modeling performed similarly on the validation set and even better on the test set than the up-to-date LSTM/ResNet architecture. For comparison with the state of the art, we also give the performances achieved by some official submissions to the MediaEval Predicting Content Interestingness task. Note that our new architecture performed better than the random baseline (each sample is randomly classified in one of the two classes) and similarly to the 3rd best official system out of 28 submissions[4].

4. DISCUSSION

From our experiments, we may draw several insights. First, as highlighted by the results presented in Table 1 and Table 2, multimodal systems were able to perform better than monomodal ones. The only two cases where multimodal systems performed worse were: 1/ when using LSTM-based systems on the MediaEval test set and 2/ when using our new n-RNN-based modeling without upsampling. In the first case, it can be explained by the low generalization capability of the multimodal system, due in part to the small size of the dataset. The size issue coupled with the unbalance of the dataset when no upsampling is applied is in turn one potential explanation of the second case.

Our results on the two datasets also differ substantially. We envision several reasons for this matter of fact. First, this may show once again that we encountered a generalization issue for the MediaEval dataset, due to either the size of the dataset, or to the concept of content-driven interestingness being much more difficult to infer than the social-driven interestingness. This leads to the hypothesis that both concepts differ significantly. This was confirmed by some transfer learning experiment we conducted (see results in Table 2), in which we used the DNN-based multimodal model learned on Jiang's dataset to infer interestingness on the MediaEval dataset. The low performance, *i.e.,* MAP = 0.1411, tends to prove this conceptual difference. Nevertheless this may also be partly due to the difference of data type, *i.e.,* mostly user-generated for the first case vs. professional content for the second case. Another reason for the overall low performance on the MediaEval dataset might also be the quality of the dataset and its annotations, which contains a large part of small and blurred shots. This raises also the question of the subjectivity of interestingness and the difficulty to assess content-based video interestingness from users.

Nevertheless, the best MAP value obtained by our new n-RNN-based structure is significantly higher than the baseline, leading us to the conclusion that the system indeed learns the interestingness concept, eventhough the dataset sizes might not be enough to train complex DNN architectures. Overall, our new n-RNN-based structure, by taking into account several temporal samples at the same time, offers better performances than state-of-the-art DNN architectures based on LSTM and Resnet. This shows the potential of our approach for other tasks where temporal modeling of the data is essential.

5. CONCLUSION

In this paper, we propose a generic computational model for predicting interestingness of video content, based on up-to-date deep learning architectures. We investigate its effectiveness on two video datasets, respectively with social-driven or content-driven interestingness annotations. We confirm that multimodal-based approaches, with mid-level fusion of audio and visual features, outperform monomodal-based systems for both datasets. We conclude that the two concepts *social interestingness* and *content interestingness* differ substantially. Our future work would be devoted to collect a larger dataset with reliable annotation, to understand better the intrinsic interestingness of a video. We will also orient our research toward the modeling of contextual interestingness, to better take into account the subjectivity of the notion.

[4] http://www.slideshare.net/multimediaeval/2016-mediaeval-interestingness-task-overview?qid=109a4620-f4bb-4b10-ba3b-2ffb51b52bd7&v=&b=&from_search=7

6. 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 assessment 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 remembering, despite lower accuracy for contextual details," *Psychology Association*, 2011.

[4] L. Mai and G. Schoeller, "Emotions, attitudes and memorability 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 Conference 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 Lluís 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] Christel Chamaret, Claire-Hélène Demarty, Vincent Demoulin, and Gwenalle Marquant, "Experiencing the interestingness concept within and between pictures," in *Proceeding of SPIE*, 2016, Human Vision and Electronic Imaging.

[13] 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.

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

[15] 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.

[16] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool, "Visual interestingness in image sequences," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, pp. 1017–1026.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "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.

[18] 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.

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

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

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

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