EURECOM @MediaEval 2017: Media Genre Inference for Predicting Media Interestingness

Olfa Ben-Ahmed, Jonas Wacker, Alessandro Gaballo, Benoit Huet
EURECOM, Sophia Antipolis, France
olfa.ben-ahmed@eurecom.fr,jonas.wacker@eurecom.fr,alessandro.gaballo@eurecom.fr,benoit.huet@eurecom.fr

ABSTRACT

In this paper, we present EURECOM's approach to address the *MediaEval 2017 Predicting Media Interestingness Task*. We developed models for both the image and video subtasks. In particular, we investigate the usage of media genre information (i.e., drama, horror, etc.) to predict interestingness. Our approach is related to the affective impact of media content and is shown to be effective in predicting interestingness for both video shots and key-frames.

1 INTRODUCTION

Multimedia interestingness prediction aims to automatically analyze media data and identify the most attractive content. Previous works have been focused on predicting media interestingness directly from the multimedia content [3, 6–8]. However, media interestingness prediction is still an open challenge in the computer vision community [4, 5] due to the gap between low-level perceptual features and high-level human perception of the data.

Recent research proved that perceived interestingness is highly correlated with data emotional content [9, 14]. Indeed, humans may prefer "affective decisions" to find interesting content because emotional factors directly reflect the viewer's attention. Hence, an affective representation of video content will be useful for identifying the most important parts in a movie. In this work, we hypothesize that the emotional impact of the movie genre can be a factor for the perceived interestingness of a video for a given viewer. Therefore, we adopt a mid-level representation based on video genre recognition. We propose to represent each sample as a distribution over genres (action, drama, horror, romance, sci-fi). For instance, a high confidence for the horror label inside the shot genre distribution could be perceived as more emotional (scary in this case). Therefore, this shot might be more characteristic and therefore more interesting than a neutral genre that could appear in any shot.

The media interestingness challenge is organized at MediaEval 2017. The task consists of two subtasks for the prediction of image and video interestingness respectively. The first one involves predicting the most interesting key frames. The second one involves the automatic prediction of interestingness for different shots in a trailer. For more details about the task description, related dataset and experimental setting, we refer the reader to the task overview paper [2]. The rest of the paper is organized as follows: Section 2 describes our proposed method, Section 3 presents experiments and results and finally Section 4 concludes the work and gives some perspectives.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2 METHOD

Extracting genre information from movie scenes results in an intermediate representation that may be quite useful for further classification tasks. In this section, we briefly present our method for media interestingness prediction. Figure 1 gives a brief overview over the entire framework. At first, we extract deep visual and acoustic features for each shot. We then obtain a genre prediction for each modality to finally use this prediction for the training of an interestingness classifier.

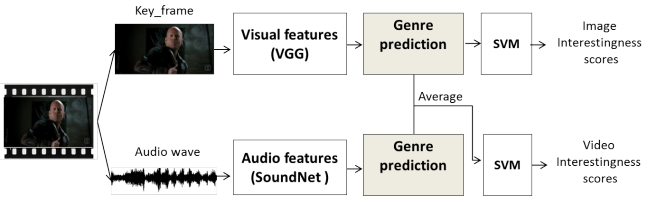

Figure 1: Framework of the proposed interestingness prediction method

2.1 Media Genre representation

The genre prediction model is based on audio-visual deep features. Using these features, we trained two genre classifiers, a Deep Neural Network (DNN) on deep visual features and an SVM on deep acoustic features.

The dataset [12] used to train our genre model contains originally 4 different movie genres: action, drama, horror and romance. We extended the dataset with an additional genre to obtain a more sophisticated genre representation for each movie trailer shot. Our final dataset comprises 415 movie trailers of 5 genres (69 trailers for *action*, 95 for *drama*, 99 for *horror*, 80 for *romance* and 72 for *sci-fi*). Each movie trailer is segmented into visual shots using the PySceneDetect tool[1]. The visual shots are automatically obtained by comparing HSV histograms of consecutive video frame (a high histogram distance results in a shot boundary). We also segment each video into audio shots using the *OpenSmile Voice Activity Detection* tool[2]. The tool automatically determines speaker cues in the audio stream which we use as acoustic shot boundaries. In total, we trained our two genre predictor models on 29151 visual and on 26144 audio shots. The visual shots are represented by key-frames. We select the middle frame in a shot as a key-frame. Visual features are extracted from these key-frames using a pretrained VGG-16 network [11]. By removing the last 2 layers, the output results in a 4096-dimensional feature vector for each keyframe. This single feature vector represents the visual information that we obtained for each shot/key-frame.

[1] http://pyscenedetect.readthedocs.io/en/latest/
[2] https://github.com/naxingyu/opensmile/tree/master/scripts/vad

Figure 2: DNN architecture for key-frame genre prediction.

2.1.1 Visual feature learning. We use the DNN architecture proposed by [12] to make genre predictions on visual features. The architecture is shown in Figure 2. The Dropout regularization is used to avoid overfitting and to optimize training performance. The output is squashed into a probability vector over the 5 genres using Softmax. We use mini-batch stochastic gradient descent on a batch size of 32 to train the network. Categorical cross entropy is used as a loss function and we train the network over 50 epochs.

2.1.2 Acoustic feature learning. Adding the audio information surely plays an important role for content analysis in videos. Most of the approaches in related work only focus on hand-crafted audio features such as the Mel Frequency Cepstrum Coefficients (MFCC) or spectrograms, with either traditional or deep classifiers. However, those audio features are rather low-level representations and are not designed for semantic video analysis. Instead of using such classical audio features, we extract deep audio features from a pre-trained model called Soundnet [1]. The latter has been learned by transferring knowledge from vision to sound to ultimately recognize objects and scenes in sound data. According to the work of Ayter *et al.* [1], an audio feature representation using Soundnet reaches state-of-the-art accuracy on three standard acoustic scene classification datasets. In our work, features are extracted from the fifth convolutional layer of the 8-layers version of the Soundnet model. For the training on audio features, we used a probabilistic SVM with a linear kernel and a regularization value of $C = 1.0$.

2.2 Interestingness classification

Our genre model can be used for both the image and video subtasks. Indeed, we train two separate genre classifiers (i.e., one based on audio and one based on visual features). Therefore, we end up with two probability vector outputs for respectively the visual and audio inputs. In order to obtain the final genre distribution for the video shots, we simply take the mean of both probability vectors. This probabilistic genre distribution is our mid-level representation and thus serves as the input for the actual interestingness classifier. A Support Vectors Machine (SVM) binary classifier is then trained on these features to predict with a confidence score whether a shot/image is considered interesting or not. For the video subtask, we also performed experiments using only the visual information of the video shots. For this we used the genre prediction model based on the extracted VGG features from the video key-frames. To evaluate the performance of our interestingness model, we tested several SVM kernels (linear, RBF and sigmoid) with different parameters on the development dataset. A high number of experiments with a grid search in order to optimize kernel parameters tended to classify almost all the samples as non interesting. This may be due to the imbalanced labels of the training data. Hence, we opted for a weighted version of SVM classification where the minority class receives a higher misclassification penalty. We also take into account the confidence scores of the development set samples during training by giving a larger penalty to samples with high confidence scores, and a small penalty to samples with low confidence scores.

3 EXPERIMENTS AND RESULTS

The evaluation results of our models on the test data provided by the organizers are shown below. We submitted two runs for the image classification and five for the video classification task. Table 1 reports the MAP and the MAP@10 scores for our various model configurations returned by the task organizers.

Task	Run	Classifier	MAP	MAP@10
Image	1	SVM - Sigmoid kernel	**0.2029**	**0.0587**
	2	SVM - Linear kernel	0.2016	0.0579
Video	1	Sigmoid kernel: gamma=0.5, C=100	0.2034	0.0717
	2	Polyn. kernel: degree=3	0.1960	0.0732
	3	Polyn. kernel: degree=2	0.1964	0.0640
	4	Sigmoid kernel: gamma=0.2, C=100	**0.2094**	**0.0827**
	5	Sigmoid kernel: gamma=0.3 , C=100	0.2002	0.0774

Table 1: Official evaluation results on test data

For the image subtask, the MAP values are quite similar for both linear and sigmoid SVM kernels. For the video subtask, decent results in MAP values are already achieved with visual key-frame classification (run 2 and 3). When using both modalities (run 1, 4 and 5), averaging audio and video genre predictions, results show a slight performance gain. However, we obtain a larger improvement when looking at the MAP@10 scores. Here, employing both modalities outperforms the pure key-frame classification. Overall, an SVM with a sigmoid kernel seems more effective for the audio-visual submission than using a linear or polynomial kernel. Yet, we have only looked at SVM models in our experiments. Further improvements could be done by trying out different models as it has been done in related work [10, 13, 15]. Also, it would be interesting to apply genre prediction on all/multiple shot frames instead of employing a single key-frame. In general, we have shown that our approach is capable of making useful scene suggestions even if we do not consider it ready for commercial use yet.

4 CONCLUSION

In this paper, we presented a framework for predicting image and video interestingness that includes a genre recognition system as a mid-level representation for the data. Our best results on the testset were 20.29 and 20.94 of MAP for respectively the image and video subtasks. Obtained results are promising especially for the video subtask. Future works include the joint learning of audio-visual features and the integration of temporal information to describe the evolution of audio-visual features over video frames.

ACKNOWLEDGMENTS

The research leading to this paper was partially supported by Bpifrance within the NexGenTV Project (F1504054U). The Titan Xp used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems*. 892–900.

[2] Claire-Hélène Demarty, Mats Viktor Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc QK Duong, Frédéric Lefebvre, and others. Media interestingness at Mediaeval 2017. In *Proceedings of MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017*.

[3] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. 2014. Interestingness prediction by robust learning to rank. In *Proceedings of the European Conference on Computer Vision*. Zurich, Switzerland, September 6-12, 488–503.

[4] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, December 1-8, 2013*. 1633–1640.

[5] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, December 11-18, 2015*. 3090–3098.

[6] Michael Gygli and Mohammad Soleymani. 2016. Analyzing and Predicting GIF Interestingness. In *Proceedings of ACM Multimedia, Amsterdam, The Netherlands, October 15-19, 2016*. New York, NY, USA, 122–126.

[7] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and Predicting Interestingness of Videos. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, July 14-18, 2013*.

[8] Yang Liu, Zhonglei Gu, Yiu-ming Cheung, and Kien A. Hua. 2017. Multi-view Manifold Learning for Media Interestingness Prediction. In *Proceedings of ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, June 6-9, 2017*. New York, NY, USA, 308–314.

[9] Soheil Rayatdoost and Mohammad Soleymani. 2016. Ranking Images and Videos on Visual Interestingness by Visual Sentiment Features. In *Proceedings of the MediaEval 2016 Workshop, Hilversum, Netherlands, October 20-21, 2016*.

[10] G. S. Simoes, J. Wehrmann, R. C. Barros, and D. D. Ruiz. 2016. Movie genre classification with Convolutional Neural Networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 259–266.

[11] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition, Technical report. *CoRR* abs/1409.1556 (2014).

[12] K S Sivaraman and Gautam Somappa. 2016. MovieScope: Movie trailer classification using Deep Neural Networks. *University of Virginia* (2016).

[13] John R. Smith, Dhiraj Joshi, Benoit Huet, Hsu Winston, and Jozef Cota. 2017. Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation. In *Proceedings of ACM Multimedia*. October 23-27, 2017, Mountain View, CA, USA.

[14] Mohammad Soleymani. 2015. The Quest for Visual Interest. In *Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, October 26-30, 2015*. New York, NY, USA, 919–922.

[15] Sejong Yoon and Vladimir Pavlovic. 2014. Sentiment Flow for Video Interestingness Prediction. In *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia (HuEvent '14)*. New York, NY, USA, 29–34.