The MediaEval 2018 Movie Recommendation Task: Recommending Movies Using Content

Yashar Deldjoo[1], Mihai Gabriel Constantin[2], Athanasios Dritsas[3],
Bogdan Ionescu[2], Markus Schedl[4]
[1]Politecnico di Milano, Italy, [2]University Politehnica of Bucharest, Romania, [3]Delft University of Technology, Netherlands, [4]Johannes Kepler University Linz, Austria
deldjooy@acm.org,mgconstantin@imag.pub.ro,a.dritsas@student.tudelft.nl
bionescu@imag.pub.ro,markus.schedl@jku.at

ABSTRACT

In this paper we introduce the MediaEval 2018 task Recommending Movies Using Content. It focuses on predicting overall scores that users give to movies, i.e., average rating (representing overall appreciation of the movies by the viewers) and the rating variance/standard deviation (representing agreement/disagreement between users) using audio, visual and textual features derived from selected movie scenes. We release a dataset of movie clips consisting of 7K clips for 800 unique movies. In the paper, we present the challenge, the dataset and ground truth creation, the evaluation protocol and the requested runs.

KEYWORDS

movie rating prediction, movie recommender systems, multimedia features, audio, visual, textual descriptors, clips, trailers

1 INTRODUCTION

A dramatic rise in the generation of video content has been witnessed in recent years. Video recommender systems (RS), play an important role in helping users of online streaming services to cope with the information overload. Video recommendation systems are traditionally powered by either collaborative filtering (CF) models which leverage the correlations between users' consumption patterns or content-based filtering (CBF) approaches typically based on *textual metadata*, either editorial, e.g., genre, cast, director, or user generated e.g., tags, reviews [1, 15].

The goal of the *MediaEval Movie Recommendation Task* is to use content-based audio, visual and metadata features and their multimodal combinations to *predict* how a movie will be received by its viewers by predicting global ratings of users and the standard deviation of ratings [7]. The task uses as input movie clips instead of the full-length movies, which makes it more versatile and effective as clips are more easily available than the full movies. There are two main useful outcomes of this task: firstly, by predicting the average ratings that users give to movies, such techniques can be exploited by producers and investors to decide whether or not to adopt the production of similar movies; secondly and **more importantly** the task is laying the groundwork for CBF movie recommendation where recommendations are tailored to match the individual preferences of users on the audio-visual content and the descriptive metadata. As for the latter, the current MediaEval task looks into predicting the variance of the ratings whose correct predictions imply the ability of the prediction system to differentiate between the preferences of different users or groups of users which can be exploited by current CBF movie recommender systems. In contrast to the de facto CF approach widely adopted by the community of RS, the CBF approach can handle the cold-start problem for items where the newly added items lack enough interactions (impeding the usability of CF approach) and can also help systems respect user privacy [3, 4]. This paper presents an overview of the task, the features provided by the organizers, a description of the ground truth and evaluation methods as well as of the required runs.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2 TASK DESCRIPTION

The goal of the task is to use content-based features to predict how a movie is received by its viewers. Task participants must create an automatic system that can predict the *average ratings* that users will assign to movies (representing the overall appreciation of the movie by the audience) and also the *rating variance* (representing the agreement/disagreements between user ratings)[1]. The input to the system is a set of audio, visual, and text features derived from selected movie scenes (movie clips).

The novelty of this task is that it uses *movie clips* instead of movie trailers as chosen by most of previous works both in the multimedia and recommendation fields [4, 6, 11]. Movie trailers for the most part are free samples of a film that are packaged to communicate a feeling of the movie's story. Their main goal is to convince the audience to come back for more when the film opens in theaters. For this reason, the trailers are usually made with lots of thrills and chills. Movie clips, however, focus on a particular scene and display the scene at the natural pace of the movie . The two media types communicate different information to their viewers and can evoke different emotions [14] which in turn strongly effect the users' perception and appreciation, i.e. ratings, of the movie. To give an example, compare from the movie "Beautiful Girls" (1996) the official trailer,[2] a movie clip (A girl named Marty),[3] and another movie clip (Ice skating with Marty)[4] all taken from the same movie.

3 DATA

Participants are supplied with audio and visual features extracted from movie clips as well as associated metadata (genre and tag

[1]Note that in fact it is required to predict standard deviation of ratings, cf. Section 5 but due to intelligibility we use the term "variance" instead of standard deviation.
[2]https://www.youtube.com/watch?v=yfQ5ONwWxI8
[3]https://www.youtube.com/watch?v=4K8M2EVnoKc
[4]https://www.youtube.com/watch?v=M-h1ERyxbQ0

labels). These content features resemble the content features of our recently released movie trailer dataset MMTF-14K [4, 5]. However, unlike in MMTF-14K, in the movie clips dataset used in the MediaEval task at hand, each movie can be associated with several clips.

The development dataset provides features computed from 6877 clips corresponding to 796 unique movies from the well-known MovieLens 20M dataset (ml-20m) [10]. The task makes use of the user ratings from the ml-20m dataset in order to calculate the grountruth, namely the per-movie global average rating and rating variance. The YouTube IDs of the clips are also available in the movie names of the clips. For example, 000000094_*2Vam2a4r9vo* represents a clip in the dataset with the ml-ID 94 and the YouTube ID *2Vam2a4r9vo*[5]. Each movie has on average $\frac{6877}{796} = 8.63$ associated clips where this value is calculated over both the training and test set. The content descriptors are organized in three categories described next.

3.1 Metadata

The *metadata descriptors* (found in the folder named Metadata) are provided as two CSV files containing *genre* and *user-generated tag* features associated with each movie. The metadata features come in pre-computed numerical format instead of the original textual format for ease of use. The metadata descriptors are exactly the same as with our MMTF-14K trailer dataset [4, 5].

3.2 Audio features

The *Audio descriptors* (found in the folder named Audio) are contained in two sub-folders: block level features (BLF) [17] and i-vector features [8, 16, 17]. The BLF data includes the raw features of the 6 sub-components (sub-features) that describe various audio aspects: spectral aspects (spectral pattern, delta spectral pattern, variance delta spectral pattern), harmonic aspects (correlation pattern), rhythmic aspects (logarithmic fluctuation pattern), and tonal aspects (spectral contrast pattern). The i-vector features, describing timbre, include different parameters for Gaussian mixture models (GMM) equal to (16, 32, 64, 256, 512), the total variability dimension (tvDim) equal to (10, 20, 40, 200, 400). The Block level features folder has two subfolders: "All" and "Component6"; the former contains the super-vector created by concatenating all 6 sub-components, the latter contains the raw feature vectors of the sub-components in separate CSV files. The i-vector features folder contains individual CSV files for each of the possible combinations of the two parameters GMM, and tvDim.

3.3 Visual features

The *Visual descriptors* (found it the folder named Visual) are contained in two sub-folders: Aesthetic visual features [9, 13] and Deep AlexNet Fc7 features [2, 12], each of them including different aggregation and fusion schemes for the two types of visual features. These two features are aggregated by using four basic statistical methods, each corresponding to a different sub-folder, that compute a video-level feature vector from frame-level vectors by using: average value across all frames (denoted "Avg"), average value and variance ("AvgVar"), median values ("Med") and finally median and median and median absolute deviation ("MedMad"). Each of the four aggregation sub-folders of the Aesthetic visual features folder contains CSV files for three types of fusion methods: early fusion of all the components (denoted All), early fusion of components according to their type (color based components denoted Type3Color, object based components - Type3Object and texture - Type3Texture) and finally each of the 26 individual components with no early fusion scheme (example: the colorfulness component denoted Feat26Colorfulness), therefore resulting in a total of 30 files in each sub-folder. Regarding the AlexNet features, in our context, we use the output values extracted from the fc7 layer. For this reason, no supplementary early fusion scheme is required or possible, and only one CSV file is present inside each of the four aggregation folders.

[5]https://www.youtube.com/watch?v=2Vam2a4r9vo

4 RUN DESCRIPTION

Every team can submit up to 4 runs, 2 runs for prediction score for rating average and 2 runs for rating std. For each score type, the first run is expected to contain the prediction score for the best unimodal approach (using visual information, audio or metadata) and the second run, hybrid approach that consider all modalities. Note that in all these runs, the teams should think how to temporally aggregate clip-level information into movie-level information (each movie on average is assigned 8 clips). This task is novel in two regards. First, the dataset includes movie clips instead of trailers, thereby providing a wider variety of the movie's aspects by showing different kinds of scenes. Second, including information about the ratings' variance allows to assess users' agreement and to uncover polarizing movies.

5 GROUND TRUTH AND EVALUATION

The evaluation of participants' runs is realized by predicting users' overall ratings for which we use the standard error metric root-mean-square-error (RMSE) between the predicted scores and the actual scores according to the ground truth (as given in the MovieLens 20M dataset), $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2}$ where N is the number of scores in the test set on which the system is validated, s_i is the actual score of users given to item i and $\hat{s}_i$ is the predicted score. Two types of scores are considered for evaluation

(1) average ratings
(2) standard deviation of ratings

The standard deviation of ratings is chosen to measure the agreement/disagreements between user ratings thereby building the groundwork for personalized recommendation. It should be reminded that during test data release, participants are provided only with the IDs of test movie clips where they are expected to predict both of the above scores.

6 CONCLUSIONS

The 2018 Movie Recommendation Task provides an unified framework for evaluating participants' approaches to the prediction of movie ratings through the usage of movie clips and audio, visual and metadata features and their hybrid combinations. Details regarding the methods and results of each individual run can be found in the working note papers of the MediaEval 2018 workshop proceedings.

REFERENCES

[1] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender systems*. Springer, 139–166.

[2] Mihai Gabriel Constantin and Bogdan Ionescu. 2017. Content description for Predicting image Interestingness. In *Signals, Circuits and Systems (ISSCS), 2017 International Symposium on*. IEEE, 1–4.

[3] Yashar Deldjoo. 2018. *Video recommendation by exploiting the multimedia content.* Ph.D. Dissertation. Italy.

[4] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2018. Audio-Visual Encoding of Multimedia Content to Enhance Movie Recommendations. In *Proceedings of the Twelfth ACM Conference on Recommender Systems*. ACM. https://doi.org/10.1145/3240323.3240407

[5] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. MMTF-14K: A Multifaceted Movie Trailer Dataset for Recommendation and Retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys 2018)*. Amsterdam, the Netherlands.

[6] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. 2018. Using Visual Features based on MPEG-7 and Deep Learning for Movie Recommendation. *International Journal of Multimedia Information Retrieval* (2018), 1–13.

[7] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2018. Content-Based Multimedia Recommendation Systems: Definition and Application Domains. In *Proceedings of the 9th Italian Information Retrieval Workshop (IIR 2018)*. Rome, Italy.

[8] Hamid Eghbal-Zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. 2015. I-Vectors for Timbre-Based Music Similarity and Music Artist Classification.. In *ISMIR*. 554–560.

[9] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.

[10] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2016), 19.

[11] Yimin Hou, Ting Xiao, Shu Zhang, Xi Jiang, Xiang Li, Xintao Hu, Junwei Han, Lei Guo, L Stephen Miller, Richard Neupert, and others. 2016. Predicting movie trailer viewer's "like/dislike" via learned shot editing patterns. *IEEE Transactions on Affective Computing* 7, 1 (2016), 29–44.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[13] Congcong Li and Tsuhan Chen. 2009. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing* 3, 2 (2009), 236–252.

[14] Robert Marich. 2013. *Marketing to moviegoers: a handbook of strategies and tactics.* SIU Press.

[15] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34.

[16] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *IJMIR* 7, 2 (2018), 95–116. https://doi.org/10.1007/s13735-018-0154-2

[17] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner. 2011. A Refined Block-level Feature Set for Classification, Similarity and Tag Prediction. In *7th Annual Music Information Retrieval Evaluation eXchange (MIREX 2011)*. Miami, FL, USA.