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, The 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 global scores of users given to movies, i.e., average rating (representing global appreciation of the movies by the viewers) and the rating variance (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, content-based, audio descriptors, visual descriptors, textual descriptors, multimodal fusion, movie clips, movie trailers, temporal aggregation

1 INTRODUCTION

The media and entertainment (M&E) industry is a several-hundred-billion-dollar industry. The global revenue obtained from the M&E market in 2017 was approximately 2 trillion dollars. Specifically the movie industry occupies an important part of the business market in the M&E industry, in addition to its vast cultural and sociological impacts [1, 2]. Producing a new movie means that the company is betting on this movie's success. Often this success or failure is determined in the first weekend of play. With a specific end goal to make this opening successful, producers and investors must utilize various professional promotional methodologies, including movie trailers and movie clips, to publicize the film for a long time frame before its release. An indicator of the movie's success is the ratings users assign to it, for example the ratings on IMDB[1]. These ratings are a major factor in determining whether users watching the movie have liked it or not. Being able to accurately predict such ratings means being able to predict into the future whether a movie will be successful or not. For this purpose, it is necessary to develop machine learning techniques that can predict the success of a movie. Used in the hand of producers and investors, such techniques are vital to decide whether or not to adopt the production of similar movies.

In the literature, most of considered factors to the success of movies have been focused on pre-release information such as metadata including actors' names, writers, producers, genre, production company, and so forth. A wealth of information in the actual video signals is thereby disregarded. In order to fill this gap, the goal of the *MediaEval Movie Recommendation Task* is to use content-based features to *predict* how a movie will be received by its viewers [8].

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 global appreciation of the movie by the audience) and also the *rating variance* (representing the agreement/disagreements between user ratings)[2]. 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 [3, 4, 6, 7, 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. Their goal is to elicit particular emotions in the audience. The two media types can evoke different emotions in their viewers [14]. To give an example, compare from the movie "Beautiful Girls" (1996) the official trailer,[3] a movie clip (A girl named Marty),[4] and another movie clip (Ice skating with Marty).[5]

The task has two objectives: one is to explore ways in which multimedia content is useful for movie recommendation; the other one is to investigate in which ways information coming from different parts of a movie can be analyzed for recommendation.

[1]https://www.imdb.com

[2]Note that standard deviation of ratings is in fact required to predict, refer to section 5.
[3]https://www.youtube.com/watch?v=yfQ5ONwWxI8
[4]https://www.youtube.com/watch?v=4K8M2EVnoKc
[5]https://www.youtube.com/watch?v=M-h1ERyxbQ0

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

3 DATA

Participants are supplied with audio and visual features extracted from movie clips as well as associated metadata (genre and tag labels). These content features resemble the content features of our recently released movie trailer dataset MMTF-14K[6] [4, 5]. However, unlike in MMTF-14K, in our movie clips dataset used in the MediaEval task at hand, each movie can contain several associated clips. The movieIds correspond to the well-known MovieLens 20M dataset [10].

The development dataset provides features computed from 6877 clips corresponding to 796 unique movies from the well-known MovieLens 20M dataset. The task makes use of the user ratings from the MovieLens dataset in order to provide the global average rating, and the rating standard deviation of the movies. The YouTube IDs of the clips are also available in the moviename of the clips. To provide an example, 000000094_*2Vam2a4r9vo* represent a movie with the corresponding MovieLens Id 94 and the YouTube ID *2Vam2a4r9vo*, the video can be watched on YouTube as the link in the footnote[7]. As the numbers of clips show, each movie has $\frac{6877}{796} = 8.63$ associated clips.

The content descriptors are organized in three categories described in the following:

3.1 Metadata

The *metadata descriptors* (found at the folder named Metadata) provide 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 those for our MMTF-14K trailer dataset. More information about the textual feature can be found at [4, 5].

3.2 Audio features

The *Audio descriptors* (found at the folder named Audio) contain two sub-folders: block level features (BLF) [16] and i-vector features [15, 16]. While the BLF data includes the raw features of the 6 sub-components, the i-vector features 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 resulted from 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 at the folder named Visual) provides two sub-folders: Aesthetic visual features [9, 13] and Deep AlexNetFc7 features [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 included in 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 generating a total of 30 files in each sub-folder. Regarding the AlexNet features, in our context, we use the extracted output values of the fc7 layer, and therefore no supplementary early fusion scheme is required or possible, and therefore only one CSV file is present inside each of the four aggregation folders.

[6] https://mmprj.github.io/mtrm_dataset/index
[7] https://www.youtube.com/watch?v=2Vam2a4r9vo

4 RUN DESCRIPTION

Every team can submit up to 12 runs, 6 per each score type (i.e., rating average and rating std). run1 - using visual information only; run2: using audio information only; run3: using textual information only. The last runs, run4, run5 and run6, are general ones, i.e., any approach is allowed, e.g., hybrid approaches 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 contains 8 clips). This task is novelty compared with the same task using movie trailers instead (for each movie there exists only 1 trailer).

5 GROUND TRUTH AND EVALUATION

The representativeness of clips with respect to movies is realized by predicting users' global 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)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|s_i - \hat{s}_i|} \quad (1)$$

where N is the number of movies 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 content features are provided for all movie clips during the development data release. However, 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.

REFERENCES

[1] 2018. 2017 Top Markets Report Media and Entertainment. https://www.trade.gov/topmarkets/pdf/Top%20Markets%20Media%20and%20Entertinment%202017.pdf. (2018). Accessed: 2018-12-27.
[2] 2018. Media and Entertainment Industry Overview. https://investmentbank.com/media-and-entertainment-industry-overview/. (2018). Accessed: 2018-12-27.
[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, Paolo Cremonesi, Markus Schedl, and Massimo Quadrana. 2017. The effect of different video summarization models on the quality of video recommendation based on low-level visual features. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 20.

[7] 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.

[8] 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.

[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] 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

[16] 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.