Retrieving Relevant and Diverse Movie Clips Using the MFVCD-7K Multifaceted Video Clip Dataset

Yashar Deldjoo
Polytechnic University of Bari, Italy
yashar.deldjoo@poliba.it

Markus Schedl
Johannes Kepler University Linz, Austria
markus.schedl@jku.at

***Abstract*—Multimedia search is an emerging area in information retrieval (IR) and recommender systems (RS) research. However, there is a lack of standardized audiovisual datasets that include rich content descriptors, which are a necessity in content-based IR and RS. The contributions of this paper are twofold: First, we present a new *multimedia dataset of movie clips*, named MFVCD-7K Multifaceted Video Clip Dataset, that comes with low-level and semantic multimodal descriptions of their content (textual, audio, and visual). In addition, we showcase the use of this dataset for a novel *content-based video clip retrieval and result diversification task* we introduce. We investigate baseline algorithms for retrieval and diversification, and provide experimental results according to relevance and diversity measures. We believe that both dataset and baseline results constitute an important asset for the IR, RS, and multimedia communities.**

***Index Terms*—multimedia, recommender system, content-based filtering, movie clips, Movie Genome**

I. Introduction

Multimedia search is an emerging area in information retrieval (IR) [10] and recommender systems (RS) research [11], not least because of the ever increasing amount of user-generated multimedia content [12], [13]. However, there exist only few public multimedia datasets that can be used for content-based video retrieval and recommendation. Most of them lack audiovisual content features or descriptors. Against this background, we introduce a novel multimedia dataset (MFVCD-7K) of video clips together with a rich set of content features: low-level and semantic descriptors (textual, audio, visual). The dataset includes *video clips*, i.e., selections of movie parts without any manual editing, in contrast to trailers or full movies, as trailers might not be representative of the movie and full movies are usually not freely available. Furthermore, we propose a novel video retrieval and diversification task: based on a query movie, retrieve relevant and diverse video clips of related movies. We present results achieved by several baseline algorithms on the MFVCD-7K dataset.

II. Related Work

Important movie and video datasets used in IR and RS research are summarized in Table I. The most frequently used ones originate from movie content providers or reviewing platforms, such as MovieLens[1] or the Internet Movie Database (IMDB).[2] Datasets made public by these companies commonly include metadata about movies and preference information of users, which has enabled research on personalization, retrieval, and recommendation using real-world data.

The *MovieLens* (ML) datasets provided by GroupLens are perhaps the most commonly adopted ones in the RS community [1]. They come in different versions (e.g., ML-100K, ML-1M, ML-10M, and ML-20M), which for the most part differ in terms of number of users and items. While earlier versions (ML-100K, ML-1M) provide user demographics (e.g., age and gender), later versions include user-generated tags instead.

The dataset *Rotten Tomatoes Movie Reviews* [2] provides reviews (e.g., critics' reviews, critics' ratings, percentage of favorable reviews) and metadata (e.g., genre, director, writer) for about 1.5K movies. In addition, this dataset includes users' overall ratings on movies and a number of descriptive metadata such as box office earning, and movie synopsis.

The *IMDB Movie Dataset* [3] provides information about 14.7K movies, gathered from IMDB and preprocessed to facilitate research on machine learning tasks. The metadata includes genre, year, duration, number of awards, average ratings and rating count. The *IMDB Movie Reviews* dataset [14] has been created to serve as benchmark for sentiment classification. The dataset comprises about 50K reviews for 7.1K movies and sentiment polarity annotations (positive/negative).

The *Yahoo! Movies Webscope* dataset [5] is another related dataset that supplies a small percentage of user ratings on 11.9K movies and provides 211.2K reviews. The dataset also includes some descriptive information such as cast, synopsis, genre, average ratings, and awards. However, the dataset is limited to movies released prior to November 2003.

The *LDOS-CoMoDa* [6] dataset contains community ratings given to movies as well as 12 pieces of contextual information e.g., time, day, season, weather, mood and health condition, facilitating research on context-aware movie RS.

The *Anime* dataset [7] contains information on users' individual preferences (explicit ratings and whether the user watched the movie) for about 12.3K Anime movies collected from MyAnimeList.net.[3] It also includes descriptive metadata (e.g., genre, episode, or number of community members).

[1] https://www.movielens.org

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

[3] https://www.myanimelist.net

978-1-7281-4673-7/19/$31.00 ©2019 IEEE

TABLE I
MOST RELEVANT MOVIE/VIDEO DATASETS USED IN IR AND RS RESEARCH. COLUMN "CONTENT FEATS." INDICATES THE KIND OF DESCRIPTORS PROVIDED: M - METADATA, A - AUDIO, AND V - VIDEO. COLUMN "VDL" INDICATES WHETHER THE DATASET INCLUDES DOWNLOAD LINKS TO THE ACTUAL VIDEO CONTENT.

Dataset	Video Type	No. Videos	Content Feats.	Additional Data (Selection)	VDL
MovieLens 20M (ML-20M) [1]	movies	26.7K	M	ratings, tags, genre, year	✓
Rotten Tomatoes Movie Reviews [2]	movies	1.5K	M	average rating, reviews, ratings, cast, box office	✗
IMDB Movie Dataset [3]	movies	14.7K	M	average rating, rating count, genre, year, awards	✗
IMDB Movie Reviews [4]	movies	7.1K	M	reviews, review sentiment annotation	✗
Yahoo! Movies Webscope [5]	movies	9.1K	M	ratings, genre, cast, synopsis, awards	✗
LDOS-CoMoDa [6]	movies	1.0K	M	ratings, context (e.g., time, season, weather)	✗
Anime Database [7]	Animes	12.3K	M	ratings, genre, episode, fans	✗
LIRIS-ACCEDE [8]	video clips	9.8K	M (A, V)	valence and arousal annotations	✓
MMTF-14K [9]	movie trailers	13.6K	M, A, V	ratings, TF-IDFs of tags, genre, year	✓
MFVCD-7K	**movie clips**	**7.0K**	**M, A, V**	ratings, TF-IDFs of tags, genre, year	✓

The *LIRIS-ACCEDE* dataset [8] provides affective annotations for almost 10K video clips extracted from 160 movies. Both discrete and continuous valence and arousal annotations are included. In addition, thanks to its use in various MediaEval tasks,[4] extensions of the dataset provide some audiovisual content features and annotations of fear [15].

In 2018, we released the *MMTF-14K* dataset [9]. It provides descriptors for 13K Hollywood-type movie trailers and user ratings on movies that are linked to the ML-20M dataset. In particular, MMTF-14K includes metadata and state-of-the-art audio and visual descriptors as well as several benchmarking results. We used this dataset to solve several movie recommendation tasks (e.g., [16], [17]). A criticism of the MMTF-14K dataset is the underlying assumption that movie trailers are representative of full movies. Movie trailers are human-edited and artificially made with lots of thrills and chills since their main goal is to convince the audience to watch the movie. Therefore, the scenes in trailers are usually drawn from the most exciting, funny, or otherwise noteworthy parts of the film, which is a strong argument against the representativeness of trailers for the full movie. To remedy this shortcoming, we introduce a novel dataset of movie clips, named Multifaceted Video Clip Dataset (MFVCD-7K). Movie video clips focus on a particular scene and display the scene at the natural pace of the movie. Since in MFVCD-7K each movie is represented by several associated video clips, it can serve as a more *realistic* summary of the movie story than trailers.

III. THE MFVCD-7K DATASET

MFVCD-7K supplies several state-of-the-art audio and visual features as well as metadata (movie, genre, bag-of-word representations of tags, and YouTube identifiers) for all included movie clips. Each clip focuses on a particular scene in the movie with a specific semantic (e.g., a fight or a dialog). The dataset covers 6,877 clips corresponding to 796 unique movies. Hence, each movie is associated with 8.63 clips on average. All 796 movies are linked to the ML-20M dataset from which it is possible to obtain users' individual ratings to movies. The MFVCD-7K dataset can be downloaded from https://mmprj.github.io/MFVCD-7K. The following content features are provided in the dataset:

Metadata (textual) features comprise genre information and user-generated keywords (tags). The former are represented by multi-hot encoded MovieLens genres; the latter by 10,000 dimensional TF-IDF feature vectors computed from user-generated tags. In addition, YouTube identifiers are provided to be able to download the actual videos.

Audio features comprise descriptors computed within the block-level framework (BLF) for audio and music processing [18] and i-vectors [19]. The former capture spectral, harmonic, tonal, and rhythmic aspects of the audio signal and are capable of incorporating information about the temporal evolution of the signal over several seconds (i.e., at the level of audio blocks). I-vectors are aggregate models that roughly describe the timbre of a signal by creating a joint representation of audio frames (typically, a few milliseconds) from Mel frequency cepstral coefficients (MFCC), modeled by Gaussian mixture models (GMM).

Visual features are represented by aesthetic visual features (AVF) [20], [21] and deep neural network features computed using AlexNet [22], [23]. The former comprise a total of 109 features designed to quantify the aesthetic appearance of an image (related to color, intensity, content diversity, texture, and discernible objects). The latter are given by a 4,096-dimensional feature vector representation of the fc7 layer of a pretrained AlexNet neural network.

IV. RETRIEVING DIVERSE VIDEO CLIPS

The example task used to demonstrate the value of the proposed dataset is that of retrieving *relevant and diverse video clips of movies* given a movie as query. The use case is that a person knows a certain movie he or she likes and wants to retrieve scenes (clips) of similar movies (in terms of genres) but covering a wider range of movies in terms of fine-grained tag annotations. This is a meaningful task because it provides users more fine-grained results compared to retrieving full movies. Also, it offers the possibility to easily browse the (typically short) clips before deciding whether to watch the full movie. This task is similar to the diversification task in image search which has already received some attention,

[4] http://www.multimediaeval.org

e.g., [24]–[26]. However, it also differs because we have to deal with two different granularity levels: movie titles are used as query and video clips as items to retrieve. Compared to result diversification in image search, only little research on the topic has been conducted in the video domains [27], [28].

V. Experiments

A. Baseline Approaches

To provide benchmarking results of baselines, we implement a simple nearest neighbor approach that uses (combinations of) multimedia features for retrieval and a rotating shuffle approach (see below) for diversification.[5] Given a movie title as a textual query, our approach first creates an aggregate feature vector from the individual feature vectors of all clips belonging to the query movie by computing the arithmetic mean over each content feature's dimension across clips. It then identifies the movie clips (considering all movies in the catalog) closest to the query in terms of a suited distance measure (cosine for TF-IDF features, Euclidean for all audiovisual features) and retrieves them. To diversify results, we use a rotating merge shuffle approach to alternatingly select clips from different movies. For this purpose, we shuffle up to 5 movies per rotation and limit to 3 the number of clips per movie to include in the results. In our experiments, we additionally include a random baseline which randomly picks k clips, ignoring the query altogether.

B. Metrics

To measure relevance, we compute average *precision@k*, investigating k's of 1, 3, 5, 10, 20, 50, and 100. A clip of a movie m_c is relevant to a query movie m_q if the Jaccard coefficient between the set of genres assigned to m_q and the set of genres assigned to m_c is at least 0.5, i.e., $J(G(m_q), G(m_c)) \geq 0.5$.[6]

To quantify diversity, we use average *tag coverage@k* and *tag entropy@k*. Coverage measures are computed both in absolute numbers and relative to the coverage of the query movie. We measure absolute tag coverage as the number of distinct tags covered by the query results. We define relative tag coverage as the absolute tag coverage of the retrieved movies (to which the retrieved clips belong) divided by the absolute tag coverage of the query movie. Tag annotations are taken from the MovieLens Tag Genome dataset [29]. It comprises 1,128 tags and provides for each pair of movie and tag a likelihood score that estimates to which extent the tag applies to the movie. We consider a tag relevant for a movie if this score is ≥ 0.7. Tag entropy is computed as the entropy of the distribution of tag occurrences over all retrieved clips.

[5]For these experiments, we extracted i-vectors with (GMM= 128, tvDim=200), average as aggregation function for AlexNet fc7 features, and median as aggregation function for AVF (empirically determined).

[6]Note that a movie can have several genres and each clip is assigned the same genres as its main movie. The retrieved clips m_c, however, can be from the query movie m_q, but also from other movies.

TABLE II
Average precision@10, tag coverage (absolute and relative)@10, and tag entropy@10 for various feature sets (A - audio, V - video, T - tags). The row "All" corresponds to the combination of i-vectors, BLF, AlexNet, AVF, and TF-IDF.

Feature	P@10	TC(abs)@10	TC(rel)@10	TEnt@10
i-vectors (A)	0.128	87.321	5.970	4.383
BLF (A)	0.252	92.585	6.126	4.414
AlexNet (V)	0.239	84.824	5.447	4.311
AVF (V)	0.196	87.918	5.769	4.353
TF-IDF (T)	0.172	82.189	5.611	4.341
All	0.258	95.057	6.297	4.438
Random	0.140	169.333	11.153	5.001

C. Results

Table II shows the performance measures for all experiments (random and nearest neighbor approach using different feature sets) at $k = 10$ retrieved clips. In addition, Figure 1 illustrates average precision, tag coverage, and tag entropy at all investigates levels of k. Please note that we intentionally omit the plot for relative tag coverage due to space limitations. The ranking is the same as that for absolute tag coverage.

Regarding *relevance*, we observe that all features except for i-vectors beat the random baseline in terms of precision. Interestingly, TF-IDF features perform inferior to all audio-visual features but i-vectors. This underlines the importance of content-based audiovisual features beyond the mere use of standard term weights for multimedia retrieval tasks. The state-of-the-art AlexNet visual features and block-level audio features both perform very well with a slightly better performance of BLF. Concatenating all features into a single feature vector yields superior results, in particular for smaller k values.

With respect to *diversity*, the random baseline outperforms all other approaches for obvious reasons. We also clearly observe the effect of the shuffling parameter in the diversification approach, which was set to 5 movies per rotation (cf. Section V-A). Therefore, if $k = 5$, the top 5 results are all taken from different movies, which leads to a similar performance of the random baseline and the approaches that leverage content features in terms of tag coverage and entropy for $k \leq 5$. No substantial differences between the feature sets can be observed except for TF-IDF which largely performs inferior (for $k \leq 50$). Concatenating all audiovisual and textual features, we obtain highest diversification among the nearest neighbor approaches.

VI. Conclusions

We presented the feature-rich multimedia dataset MFVCD-7K of movie video clips, which includes low-level and semantic multimodal content descriptions (textual, audio, and visual). Furthermore, we introduced a novel multimedia search task, i.e., retrieving relevant and diverse movie clips given a full movie as query, for which we demonstrated the use of the MFVCD-7K dataset. We provided results of baseline algorithms using a variety of content features and combinations thereof and analyzed their performance using relevance and diversity metrics. We believe that the MFVCD-7K dataset

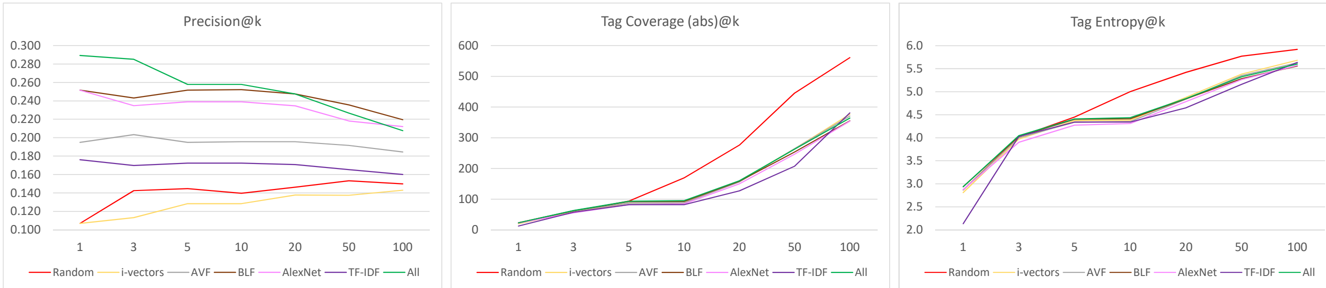

Fig. 1. Average precision, tag coverage, and tag entropy for various k values and feature sets.

represents a valuable asset not only for multimedia information retrieval but also RS research.

References

[1] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, p. 19, 2016.

[2] "Rotten Tomatoes Movie Reviews," 2018, [Accessed: 2019-02-12]. [Online]. Available: https://www.kaggle.com/rpnuser8182/rotten-tomatoes

[3] "IMDB Movies Dataset," 2016, [Accessed: 2019-02-12]. [Online]. Available: https://www.kaggle.com/orgesleka/imdbmovies

[4] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015

[5] "Yahoo! Webscope: Movies User Ratings and Descriptive Content Information, v.1.0," 2009, [Accessed: 2019-02-12]. [Online]. Available: https://webscope.sandbox.yahoo.com

[6] A. Košir, A. Odic, M. Kunaver, M. Tkalcic, and J. F. Tasic, "Database for contextual personalization," *Elektrotehniski Vestnik/Electrotechnical Review*, vol. 78, pp. 270–274, January 2011.

[7] "Anime Recommendations Database," 2016, [Accessed: 2019-02-12]. [Online]. Available: https://www.kaggle.com/CooperUnion/anime-recommendations-database

[8] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A Video Database for Affective Content Analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, January 2015.

[9] Y. Deldjoo, M. G. Constantin, B. Ionescu, M. Schedl, and P. Cremonesi, "MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018, pp. 450–455.

[10] M. S. Lew, "Multimedia Information Retrieval in the Twenty-first Century," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 1, pp. 1–2, April 2012. [Online]. Available: https://doi.org/10.1007/s13735-012-0009-1

[11] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems: Introduction and Challenges," in *Recommender systems handbook*. Springer, 2015, pp. 1–34.

[12] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Content-based multimedia recommendation systems: Definition and application domains," in *Proceedings of the 9th Italian Information Retrieval Workshop, Rome, Italy, May, 28-30, 2018.*, 2018. [Online]. Available: http://ceur-ws.org/Vol-2140/paper15.pdf

[13] Y. Deldjoo, M. Schedl, B. Hidasi, and P. Knees, "Multimedia recommender systems," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, 2018, pp. 537–538. [Online]. Available: https://doi.org/10.1145/3240323.3241620

[14] "IMDB Movie Reviews Dataset," 2016, [Accessed: 2019-02-12]. [Online]. Available: https://inclass.kaggle.com/iarunava/imdb-movie-reviews-dataset

[15] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, and M. Sjöberg, "The MediaEval 2017 Emotional Impact of Movies Task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, September 2017.

[16] Y. Deldjoo, M. G. Constantin, H. Eghbal-Zadeh, B. Ionescu, M. Schedl, and P. Cremonesi, "Audio-visual encoding of multimedia content for enhancing movie recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, 2018, pp. 455–459. [Online]. Available: https://doi.org/10.1145/3240323.3240407

[17] Y. Deldjoo, M. F. Dacrema, M. G. Constantin, H. Eghbal-Zadeh, S. Cereda, M. Schedl, B. Ionescu, and P. Cremonesi, "Movie genome: Alleviating new item cold start in movie recommendation," *User Modeling and User-Adapted Interaction (UMUAI)*, 2019.

[18] K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner, "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, October 2011.

[19] H. Eghbal-zadeh, B. Lehner, M. Schedl, and G. Widmer, "I-Vectors for Timbre-Based Music Similarity and Music Artist Classification," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, Malaga, Spain, October 2015.

[20] A. F. Haas, M. Guibert, A. Foerschner, S. Calhoun, E. George, M. Hatay, E. Dinsdale, S. A. Sandin, J. E. Smith, M. J. Vermeij *et al.*, "Can We Measure Beauty? Computational Evaluation of Coral Reef Aesthetics," *PeerJ*, vol. 3, p. e1390, 2015.

[21] C. Li and T. Chen, "Aesthetic Visual Quality Assessment of Paintings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[23] M. G. Constantin and B. Ionescu, "Content Description for Predicting Image Interestingness," in *Proceedings of the IEEE 2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, 2017, pp. 1–4.

[24] B. Yuan, X. Gao, and Z. Niu, "Discovering Latent Aspects for Diversity-Induced Image Retrieval," *IEEE MultiMedia*, vol. 25, no. 4, pp. 19–33, October 2018.

[25] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. B. D. Natale, "Multimodal Retrieval with Diversification and Relevance Feedback for Tourist Attraction Images," *ACM Transactions on Multimedia Computing and Communication Applications*, vol. 13, no. 4, pp. 49:1–49:24, August 2017. [Online]. Available: http://doi.acm.org/10.1145/3103613

[26] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller, "Result Diversification in Social Image Retrieval: A Benchmarking Framework," *Multimedia Tools and Applications*, vol. 75, no. 2, pp. 1301–1331, January 2016. [Online]. Available: https://doi.org/10.1007/s11042-014-2369-4

[27] T. Dong, S. Nishimura, and J. Liu, "Diversified and summarized video search system," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. Mountain View, CA, USA: ACM, October 2017, pp. 1263–1264. [Online]. Available: http://doi.acm.org/10.1145/3123266.3127937

[28] X. Giro-i Nieto, M. Alfaro, and F. Marques, "Diversity ranking for video retrieval from a broadcaster archive," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11. Trento, Italy: ACM, April 2011, pp. 56:1–56:8. [Online]. Available: http://doi.acm.org/10.1145/1991996.1992052

[29] J. Vig, S. Sen, and J. Riedl, "The Tag Genome: Encoding Community Knowledge to Support Novel Interaction," *ACM Transactions on Interactive and Intelligent Systems*, vol. 2, no. 3, pp. 13:1–13:44, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2362394.2362395