Audio-visual Encoding of Multimedia Content for Enhancing Movie Recommendations

Yashar Deldjoo
Politecnico di Milano
deldjooy@acm.org

Mihai Gabriel Constantin
University Politehnica of Bucharest
mgconstantin@imag.pub.ro

Hamid Eghbal-Zadeh
Johannes Kepler University
hamid.eghbal-zadeh@jku.at

Bogdan Ionescu
University Politehnica of Bucharest
bionescu@imag.pub.ro

Markus Schedl
Johannes Kepler University
markus.schedl@jku.at

Paolo Cremonesi
Politecnico di Milano
paolo.cremonesi@polimi.it

ABSTRACT

We propose a multi-modal content-based movie recommender system that replaces human-generated metadata with content descriptions automatically extracted from the visual and audio channels of a video. Content descriptors improve over traditional metadata in terms of both richness (it is possible to extract hundreds of meaningful features covering various modalities) and quality (content features are consistent across different systems and immune to human errors). Our recommender system integrates state-of-the-art aesthetic and deep visual features as well as block-level and i-vector audio features. For fusing the different modalities, we propose a rank aggregation strategy extending the Borda count approach.

We evaluate the proposed multi-modal recommender system comprehensively against metadata-based baselines. To this end, we conduct two empirical studies: (i) a system-centric study to measure the offline quality of recommendations in terms of accuracy-related and beyond-accuracy performance measures (novelty, diversity, and coverage), and (ii) a user-centric online experiment, measuring different subjective metrics, including relevance, satisfaction, and diversity. In both studies, we use a dataset of more than 4,000 movie trailers, which makes our approach versatile. Our results shed light on the accuracy and beyond-accuracy performance of audio, visual, and textual features in content-based movie recommender systems.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

recommender systems; multimedia; movies; audio features; visual features; offline evaluation; user-study

1 INTRODUCTION AND RELATED WORK

Video recordings are very complex audio-visual signals. When we watch a movie, we can effortlessly register a lot of details conveyed to us through different multimedia channels, in particular, the audio and visual modalities. As a result, movies can be described in versatile manners, which can be manifested by descriptors of visual and audio content, but also in terms of metadata, including genre or tag information.

Movie recommendation systems (RS) are traditionally powered either by collaborative filtering (CF) recommender engines or by content-based (CB) algorithms [22, 24]. CF assumes that the target user will prefer content similar to the content other like-minded users prefer. In contrast, CB approaches determine their recommendations based on similarities between the target user's preferred or consumed items and other items in the catalog, where this similarity is defined using descriptors (features) extracted from the item content, e.g., colors in an image, motion trajectories in a video, rhythm of a song, textual information on a web page.

The extent to which content-based approaches are used, even the ways "content" is interpreted, varies between domains. While in the multimedia community, extracting descriptive item features from text, audio, image, and video content is a well-researched task, the recommender systems community has considered for a long time metadata, such as title, genre, tags, actors, or plot of a movie, as the major, if not single, source for CB recommendation models, thereby disregarding the wealth of information encoded in the actual content signals [11, 12]. Addressing this research gap, we propose here a *multi-modal content-based recommender system* (CBRS) that adopts latest state-of-the-art visual and audio features in addition to metadata. [33] provides a good frame of reference on modern deep-learning based approaches on RS. Some approaches seek to create deep learning systems for recommendation using metadata information [4, 35], or single audio or visual modality [3], or applied directly on user consumption data [32, 34]. The use of audio-visual features in contrast to metadata has the additional advantage that the latter are often rare or absent for new videos, making it difficult or even impossible to provide (good) recommendations, *i.e.,* the *cold-start* problem[1] [25]. Moreover, user-generated metadata often exhibit user or community biases and might therefore not fully, or only in a distorted way, reflect the characteristics of a video [2].

Our main contributions are: (1) We propose a multi-modal CBRS for videos which uses multimedia state-of-the-art *aesthetic and deep visual features*, and *block-level and i-vectors audio features*. It outperforms the traditional use of metadata (genre and tag). To the best

[1] In RS, cold-start refers to the situation where rating and/or metadata are rare or absent. Here we use the terminology to refer to the lack of metadata.

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10…$15.00
https://doi.org/10.1145/3240323.3240407

of our knowledge, this is a novel approach, existing systems being limited either to use single modalities [7–10] or deprecated, low-level descriptors [23]; (2) We propose a rank aggregation strategy based on Borda count [1] to fuse recommendations from different (heterogeneous) sources, outperforming the results obtained with traditional Borda count; (3) Our recommendation framework uses as input movie trailers instead of the entire movies, which makes it more versatile and effective;[2] (4) We evaluate our CBRS comprehensively, via two wide and articulated empirical studies: (i) a *system-centric* experiment to measure the offline quality of recommendations in terms of accuracy-related performance and beyond-accuracy measures [17], including novelty and diversity; (ii) a *user-centric* online experiment, measuring different subjective metrics, including relevance, satisfaction, and diversity.

Our work most closely resembles [14], but improves it in several directions: (1) we exploit audio content; (2) instead of hand-crafted features of low dimensionality, we integrate state-of-the-art audio-visual features as well as temporal aggregation techniques; (3) instead of basic concatenation, we propose an extension of Borda count for hybridization; (4) for the user study, we performed decent sanity checks and removal of unreliable user input, as well as extended the questionnaire to cover the full instrument proposed by [13, 20].

2 CONTENT DESCRIPTORS

2.1 Audio Features

We investigate state-of-the-art *block-level features* and *i-vector features* used in audio and music processing tasks, including speaker identification, music classification, similarity, and recommendation [19, 27].

Block-level features (BLF) [28] are extracted from audio segments of typically a few seconds duration and can therefore capture temporal aspects of an audio recording to some degree, which contrasts frame-level features which operate on much shorter units. The BLF framework [28] defines six features that capture: 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 extraction process results in a 9,948-dimensional feature vector per video.

I-vector features [5] represent a state-of-the-art representation learning technique in audio-related domains, such as speech processing, music recommendation, and acoustic scene analysis [26, 31]. An i-vector is a fixed-length, low-dimensional representation of an acoustic signal. I-vectors are latent variables that capture, in our case, the deviation of a video clip representation from the average representation of all videos. We train a Gaussian mixture model (GMM) from frame-level Mel frequency cepstral coefficients (MFCC) to reflect the average distribution of all videos in the acoustic feature space. To capturing the shift of the adapted model of an individual video from the average representation, factor analysis is then applied, resulting in a low-dimensional latent variable vector. The results are reported using 20-dimensional MFCCs, 512 number of components in GMM and final i-vector dimensionality of 200.

[2]Trailers are more easily available than the full movies.

2.2 Visual Features

We select two types of state-of-the-art visual features: *aesthetic visual features* and *deep learning features*.

Aesthetic visual features (AVF) [15] quantify aesthetics in photographic images and art paintings. They are driven by concepts such as image composition, color theory, or interestingness. The features are grouped into three categories: color-based, texture-based, and object-based [15], adding up to a 107-dimensional vector for each frame.

Deep learning features resulting from the AlexNet [21] deep neural network have won the 2012 Image Net Large Scale Visual Recognition Competition (ILSVRC)[3] by a large margin. For our task, we use the output of the fc7 layer for each frame in the video, a method which has shown good results in many tasks, e.g., prediction of media interestingness, text-to-video translation, and emotion recognition [29]. This yields a 4,096-dimensional descriptor vector for each frame.

2.3 Metadata Features

We also use two types of metadata features to serve as baselines: movie genre (*editorial*) and tags (*user-generated*).

Genre features describe each movie by a binary 18-dimensional vector, according to the MovieLens-20M dataset [16].

Tag features are represented using TF-IDF Bag-of-Words (BoW) model after taking a series of preprocessing steps: (1) punctuation removal, (2) tokenization and lowercase conversion, (3) stop-word removal, (4) removing words with very few ($\leq$ 2) or very many ($\geq$ 15) characters, and (5) Porter stemming. The final tag feature vector is of dimensionality 10,228 per video. The dataset containing all the content descriptors named MMTF-14K [6] is freely available for rese[4].

3 HYBRID RECOMMENDATION

Recommendations are generated with a standard k-nearest neighbor approach, where the unknown preference score (i.e., rating) for user u and item i is computed as $\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{j \in N_u(i)} s_{ij}} \sum_{j \in N_u(i)} s_{ij}\, \mathbf{r}_{uj}$ where $N_u(i)$ denotes the items rated by user u most similar to item i and s_{ij} is the similarity score between items i and j (the content-centric similarity). The recommendation is realized by rating prediction for unknown user-item combinations and then ranking the predictions.

To enhance the performance of our CBRS, we propose a hybridization approach, which is based on the Borda count aggregation method already successfully used in similar applications, e.g., group recommendation [1] or hybrid music recommendation [18]. Borda count is a rank aggregation method which converts ranks of a set of voters (recommenders in our case) into scores and aggregates them per item. Formally, given a user u and the ranking position of an item i by recommenders a and b denoted as $\sigma_u^{rec_a}(i)$ and $\sigma_u^{rec_b}(i)$, respectively, the combined score of item i for user u is given by $\mathrm{CS}(u,i) = \frac{N_a - \sigma_u^{rec_a}(i)+1}{N_a} + \frac{N_b - \sigma_u^{rec_b}(i)+1}{N_b}$ where N_a and N_b denote the total numbers of items in the corresponding rankings. We propose to extend this method by integrating *weighted*

[3]http://www.image-net.org/challenges/LSVRC/
[4]https://mmprj.github.io/mtrm_dataset/index

averaging, yielding $\text{CS}(u,i) = \sum_{k=1}^{k=N} w_k \, score_k(u,i)$ where N is the number of voters, $score_k(u,i)$ is the score of each voter for a given user-item pair (the two fractions in the original Borda count method), and w_k is the importance weight assigned to each voter. The rationale is to give an importance weight to each of the voters, so if one voter provides good-quality rankings which dominate the performance, using the simple average aggregation can deteriorate the results toward the average performance of the combined voters. A weighted average, in contrast, allows to keep the advantages of the better voter and improve it with the complementary information of the other voter(s) leading to improved overall performance. We search for the best weight combinations in the range $[0, 1]$ in such a way that $w_1 + w_2 = 1$ (in case of two voters). We investigate 100 linearly spaced combinations.

4 EXPERIMENTAL RESULTS

4.1 Data and Parameter Settings

We evaluate the proposed MRS on the MovieLens-20M (ML-20M) dataset [16]. To speed up the experiments, we randomly select 3,000 users from ML-20M, each having a minimum of 50 ratings. The final dataset contains 3,000 users, 4,899 item, 212,019 ratings and a density of 0.0144 (density of original ML-20M is equal to 0.0054).

In the offline experiment, all results are computed at single cut-off value of 4 in order to make the output compatible with the results of the user study (cf. Section 4.2.2). Results are reported based on (best) average values obtained in 5-fold cross-validation. For all feature vectors cosine is used as the similarity metric except for the genre where Jaccard index is employed.

4.2 Results and Discussions

4.2.1 Experimental Study A: Offline Evaluation. In our offline experiments we use the metadata features as baselines for evaluating the performance of the audio and visual features and the hybridized multimodal combinations of descriptors. We compare the performance of different features in terms of accuracy metrics: mean reciprocal rank (MRR), mean average precision (MAP), and recall and beyond-accuracy metrics novelty, diversity, and coverage (cf. Table 1). For both sets of metrics, we use Tukey's Kramer significance test (HSD) to make all possible pairwise comparisons [30].

For the 3 accuracy measures, the unimodal descriptors i-vector (audio) and deep (visual) perform better than their counterparts BLF (audio) and AVF (visual), in some cases better than tag, the best performing baseline metadata feature. The i-vector approach shows promising results for all 3 metrics, improving the tag's results by 9% for MRR, 5% for MAP, and 13% for recall, while the deep features outperform the tag's MRR result, but only marginally with an improvement of 2%. The multimodal approaches, especially deep + tag and i-vector + tag, consistently yields better results than i-vector, the best performing unimodal approach, thus showing that audio and visual information can help improve the performance of a recommender system. The hybrid i-vector + tag recommender yields the best results for each of the 3 accuracy metrics, revealing substantial improvements over the best results achieved with the unimodal features. Indeed, the results improve by 33% for MRR, 36% for MAP, and 28% for recall when compared with i-vector, the best performing unimodal feature. The significance tests show that for MRR and recall the improvements achieved by i-vector + deep are statistically significant compared both with i-vector and/or tag features. For the other multimodal feature sets, the combination deep + tag also achieves better results than all unimodal approaches (though not significant in all cases), while i-vector + deep does not show any significant changes when compared to i-vector.

As part of the offline experiments, we also test the performance of our hybridization approach when compared to the traditional Borda count method. The three feature combinations tested perform better with regards to the accuracy measures when aggregated with our improved method. Improvements for i-vector + deep and i-vector + tag using the proposed hybridization method when compared with the traditional Borda count approach are particularly high, 11.53% and 12.56% (on average for MRR, MAP and Recall), while the deep + tag combinations show an improvement of 1.54%. Results obtained for beyond-accuracy metrics show a similar pattern

The results for beyond accuracy measures are more diverse than those for the accuracy measures. The best results for novelty are achieved with the genre feature (9.1759), for diversity with the i-vector feature (0.8744), and for coverage with deep, i-vector + deep, and deep + tag features (1.000). The significance tests we ran for these best performing features and combinations only show a significant improvement for the 3 best performing coverage runs. A common observation for these 3 metrics is that the results seem to be very close, with small differences between the best and worst performing methods: 6.07% for novelty, 14.61% for diversity, and 6.36% for coverage. These percentage differences are indeed low when compared to the ones obtained for accuracy measures: 62.71% for MRR, 60.31% for MAP, and 55.23% for recall, therefore indicating that, when selecting a well performing algorithm, there is a very small trade-off on the beyond-accuracy measures when compared to the greater advances obtained in terms of accuracy metrics.

The results of our experimental study A indicate that the addition of multimedia features to a recommender system improves the system consistently, both in a unimodal, but especially in a multimodal approach. Furthermore, the creation of a weighted system for the multimodal hybrid approaches seems to further increase the accuracy of the recommendations returned by the system.

4.2.2 Experimental Study B: User Study. Similar to [14], we design an empirical study that considers the same recommender algorithms used in the previous offline experiment, but measures the user's *perceived* quality of the recommendations, in terms of accuracy, novelty, diversity, and overall satisfaction. Here, we focus our attention solely on unimodal recommendation approaches in 3 classes: (i) metadata: *genre* and *tag*, (ii) audio: *i-vector* and *BLF*, and (iii) visual: *deep learning features* and *AVF*. The reason is to avoid overloading the user with numerous selections and as such, to be able to obtain more reliable responses from the users collectively.

Opposite to [14], in order to handle comparison of different feature sets by users, we employ a *between-subject design*, in which each user has to compare three categories of features: audio *v.s.* visual *v.s.* metadata where the feature types in each category is randomized in each session (for each user 1 out of 8 combinations is presented at each time). To avoid possible biases, the positions of the recommendation lists are randomized for each user.

Table 1: Performance *w.r.t.* MRR, MAP, and Recall. The feature set which has the highest value for each metric and significantly outperforms the tag baseline is shown in red. The results in bold are the highest (but not significantly); cutoff value=4.

	feature name	feature type	MRR	MAP	Recall	Novelty	Diversity	Coverage
unimodal	tag	metadata	0.0213	0.0057	0.0046	8.9845	0.8685	0.9700
	genre	metadata	0.0162	0.0044	0.0039	**9.1759**	0.7553	0.9383
	i-vector	audio signal	0.0233	0.0060	0.0052	8.6514	**0.8744**	0.9994
	BLF	audio signal	0.0170	0.0045	0.0038	8.6528	0.8618	0.9998
	deep	visual signal	0.0219	0.0057	0.0043	8.6397	0.8665	**1.0000**
	AVF	visual signal	0.0187	0.0049	0.0039	8.6346	0.8735	0.9994
hybrid proposed	i-vector + deep	audio + visual	0.0232	0.0061	0.0051	8.6594	0.8733	**1.0000**
	deep + tag	visual + meta	0.0250	0.0066	0.0054	8.9720	0.8671	**1.0000**
	i-vector + tag	audio + meta	**0.0310**	**0.0082**	**0.0067**	8.9608	0.8700	0.9994

A survey containing 22 questions, originally taken from [20], reflecting various aspects of the lists is used to measure the user's perception of the recommendation lists across five factors. Questions are posed differently in positive and negative manner in order to check the consistency of results. The audience is composed of users aged between 18 and 45 who have some familiarity with the use of the web and have never used the system before. The total number of recruited subjects who also completed the task was 74 (54 male, 20 female, mean age: 24.93 years, std.: 5.11 years). In order to obtain reliable responses, the user is asked to specify how many of the movies in each of the recommendation list he/she has seen. A list is only considered if the user has seen at least one movie. The final score for each category (perceived accuracy, satisfaction, *etc.*) is a linear combination of the responses (scores) given by participants to every question belonging to the category. The results reported in the user-study are the preliminary results of an ongoing study where due to small sample size, we do not report a statistical significance test.

The final results are presented in Table 2. In terms of perceived accuracy the best algorithms are tag and deep visual features of the votes with 28% and 24% of the votes. These results are in agreement with the ones obtained in the offline experiment, at least partially meaning that as a standalone feature, the proposed state-of-the-art features based on deep learning show the most promising results compared with the other audio-visual features. The surprising result is the performance of i-vector feature which is ranked much lower among the list of recommendation performances.

The results for the perceived diversification indicate the tag and BLF features perform the best with similar score of 24%. This is while both of the visual features, AVF and deep, show lower perceived diversity. These results are interesting and indicate the strength of *audio* features for *diversification* of the recommendation list in the task of video recommendation. As for novelty, here we can see a surprising effect. First, it is the visual features AVF which has the highest amount of perceived novelty gaining as much as 47% of votes, followed by tag with 24% of the votes. The other surprising effect can be seen for genre which has attracted negative scores of the users regarding the perceived novelty. This is while our offline results show the highest novelty for genre features. The i-vector still provides good performance score for the perceived novelty and this result is in agreement with that obtained in offline experiment. Finally, the user-perceived satisfaction shows highly correlated votes from the user with that obtained by perceived accuracy, signaling that the users' perception of accuracy and satisfaction are the same. The result for both dimensions show superior performance for tag and deep features in the first place while genre and i-vector features are ranked in the second and third.

Overall, these results indicate that as a stand-alone feature user-perceived quality of recommendation is higher for the tag and deep visual features. However, for beyond accuracy metrics (such as diversity and novelty) the audio and visual features based on BLF and AVF have a promising perceived effect. This suggests to use these established features in conjunction with other rich content descriptors in the design of movie recommendation systems in order to increase diversity and novelty.

Table 2: Results of the user study along with 4 tested dimensions in a real movie recommender system.

feature name	feature type	Relevance	Diversity	Novelty	Satisfaction
tag	metadata	**0.2807**	**0.2407**	0.2381	**0.2920**
genre	metadata	0.1754	0.2037	-0.0476	0.1504
i-vector	audio	0.1053	0.1667	0.1905	0.1327
BLF	audio	0.0877	**0.2407**	0.0000	0.0885
deep	visual	0.2456	0.0926	0.1429	0.2301
AVF	visual	0.1053	0.0556	**0.4762**	0.1062

5 CONCLUSION

We presented a multi-modal content-based movie recommender systems that exploits rich content descriptors. The proposed system integrates state-of-the-art multimedia features. For fusing the different modalities, we proposed a *weighted variant of the Borda rank aggregation strategy*. Evaluation was carried out on a subset of MovieLens-20M and multimedia features extracted from 4,000 movie trailers, by (i) a *system-centric study* to measure the offline quality of recommendations in terms of accuracy-related (MRR, MAP, recall) and beyond-accuracy (novelty, diversity, coverage) performance, and (ii) a *user-centric online experiment*, measuring different subjective metrics (relevance, satisfaction, diversity). Results suggest that multimedia features can provide a good alternative to metadata (which we used as baseline), with regards to both accuracy measures and beyond accuracy measures.

REFERENCES

[1] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conf. on Recommender systems*. ACM, 119–126.

[2] Òscar Celma. 2010. *Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, Berlin, Heidelberg, Germany.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 335–344.

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191–198.

[5] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2011), 788–798.

[6] 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.

[7] 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.

[8] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, and Pietro Piazzolla. 2016. Recommending movies based on mise-en-scene design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1540–1547.

[9] 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.

[10] Yashar Deldjoo, Cristina Frà, Massimo Valla, Antonio Paladini, Davide Anghileri, Mustafa Anil Tuncil, Franca Garzotta, Paolo Cremonesi, et al. 2017. Enhancing Children's Experience with Recommendation Systems. In *Workshop on Children and Recommender Systems (KidRec'17)-11th ACM Conference of Recommender Systems*.

[11] 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, Rome, Italy, May, 28-30, 2018*. http://ceur-ws.org/Vol-2140/paper15.pdf

[12] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, and Peter Kness. 2018. Multimedia Recommender Systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM. https://doi.org/10.1145/3240323.3241620

[13] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 161–168. https://doi.org/10.1145/2645710.2645737

[14] Mehdi Elahi, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Leonardo Cella, Stefano Cereda, and Paolo Cremonesi. 2017. Exploring the Semantic Gap for Movie Recommendations. In *Proceedings of the Eleventh ACM conf. on Recommender Systems*. ACM, 326–330.

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

[16] 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.

[17] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (Dec. 2016), 42 pages. https://doi.org/10.1145/2926720

[18] Marius Kaminskas, Francesco Ricci, and Markus Schedl. 2013. Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*. Hong Kong, China.

[19] Peter Knees and Markus Schedl. 2016. *Music Similarity and Retrieval: An Introduction to Audio-and Web-based Strategies*. Vol. 36. Springer.

[20] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[21] 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.

[22] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*. Springer, 73–105.

[23] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 10.

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

[25] Sujoy Roy and Sharat Chandra Guntuku. 2016. Latent factor representations for cold-start video recommendation. In *Proceedings of the 10th ACM conf. on Recommender Systems*. ACM, 99–106.

[26] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* (2018), 1–22.

[27] 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.

[28] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. 2010. Automatic Music Tag Classification based on Block-Level Features. In *Proceedings of the 7th Sound and Music Computing conf. (SMC 2010)*. Barcelona, Spain.

[29] Yuesong Shen, Claire-Hélène Demarty, and Ngoc QK Duong. 2016. Technicolor @ MediaEval 2016 Predicting Media Interestingness Task.. In *In Proc. of the MediaEval 2016 Workshop*.

[30] David J Sheskin. 2003. *Handbook of parametric and nonparametric statistical procedures*. crc Press.

[31] Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. 2017. Music Playlist Continuation by Learning from Hand-Curated Examples and Song Features: Alleviating the Cold-Start Problem for Rare and Out-of-Set Songs. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*. ACM, 46–54.

[32] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the tenth ACM international conference on web search and data mining*. ACM, 495–503.

[33] Shuai Zhang, Lina Yao, and Aixin Sun. 2017. Deep learning based recommender system: A survey and new perspectives. *arXiv preprint arXiv:1707.07435* (2017).

[34] Yin Zheng, Bangsheng Tang, Wenkui Ding, and Hanning Zhou. 2016. A Neural Autoregressive Approach to Collaborative Filtering. In *International Conference on Machine Learning*. 764–773.

[35] Yi Zuo, Jiulin Zeng, Maoguo Gong, and Licheng Jiao. 2016. Tag-aware recommender systems based on deep neural networks. *Neurocomputing* 204 (2016), 51–60.