Using Visual Features and Latent Factors for Movie Recommendation

Yashar Deldjoo
Politecnico di Milano
Via Ponzio 34/5
20133 Milan, Italy
yashar.deldjoo@polimi.it

Mehdi Elahi
Politecnico di Milano
Via Ponzio 34/5
20133 Milan, Italy
mehdi.elahi@polimi.it

Paolo Cremonesi
Politecnico di Milano
Via Ponzio 34/5
20133 Milan, Italy
paolo.cremonesi@polimi.it

ABSTRACT

Item features play an important role in movie recommender systems, where recommendations can be generated by using explicit or implicit preferences of users on attributes such as genres. Traditionally, movie features are human-generated, either editorially or by leveraging the wisdom of the crowd.

In this short paper, we present a recommender system for movies based of Factorization Machines that makes use of the *low-level visual* features extracted automatically from movies as side information. Low-level visual features – such as lighting, colors and motion – represent the design aspects of a movie and characterize its aesthetic and style.

Our experiments on a dataset of more than 13K movies show that recommendations based on low-level visual features provides almost 10 times better accuracy in comparison to genre based recommendations, in terms of various evaluation metrics.

1. INTRODUCTION

Video-on-demand applications are characterized by the large amount of new video content produced every day. As an example, hundreds of hours of video are uploaded to YouTube every minute.

Recommender Systems based solely on Collaborative Filtering (CF) fail to provide reliable recommendations, as the large number of newly produced movies have no or very few ratings. Side information about movies (e.g., genre, cast) can be exploited to help CF deal with the new-item problem [21]

A necessary prerequisites for CF with side-information is the availability of a rich set of high-level descriptive *attributes* about movies. In many cases, such information is human-generated and prone to biases or errors.

In contrast to human-generated attributes, the content of movie streams is itself a rich source of information about low-level stylistic features that can be used to provide movie recommendations. Indeed, by analyzing a movie stream content and extracting a set of informative features, a recommender system can make personalized recommendations tailored to the users' tastes. This is particularly beneficial in the new item scenario, i.e., when a new video is added to the catalogue with absolutely no attributes available [19, 12]. While this is an interesting research direction, it has received only marginal attention of the researchers in the field.

CBRecSys 2016, September 16, 2016, Boston, MA, USA.

In this paper, we show how to use low-level visual features extracted automatically from video files as input to the recommendation algorithm. We have identified a number of visual features that have shown to be very representative of the users' feelings, according to *Applied Media Aesthetics* [23]. Our features are part of the low-level visual of a movie, and are indicative of lighting, colors and motion in the movies [9].

We have performed an exhaustive evaluation by comparing the low-level visual features, w.r.t., a more traditional set of features, i.e., genre. We have used one of the state-of-the-art recommendation algorithm, i.e., Factorization Machines (FM) [18], and fed it with either set of movie features. We have computed different relevance metrics (precision, recall, and F1) over a large dataset of more than 13M ratings provided by 182K users to more than 13K movie trailers. We have used trailers (instead of full-length movies) in order to have a scalable recommender system. In early works, we have shown that low-level features extracted from trailers of movies are equivalent to the low-level features extracted from full-length movies, both in terms of feature vectors and quality of recommendations [10, 11]. We have also performed discriminative analysis, using both trailers and full-length movies, in order to better understand the effectiveness of each low-level visual feature, individually and in combination with the others, on the performance of the recommender system. The analysis have shown high similarity between the low-level features extracted from trailers and full-length movies [10].

The results of this paper shows that recommendations based on low-level visual features achieve an accuracy, almost 10 times better than the accuracy of genre-based recommendations.

This paper extends our previous work [10], where we presented some preliminary results obtained on a much smaller dataset of 160 movies, and simpler recommendation algorithm (a content-based algorithm based on cosine similarity between items).

Our work provides a number of contributions to the research area of movie recommendation:

- we propose a novel RS that automatically analyzes the content of videos and extracts a set of low-level visual features, and uses them as side information fed to Factorization Machines, in order to generate personalized recommendations for users
- we evaluate the proposed RS using a dataset of more than 13K movies, from which we extracted the low-

level visual features

- the dataset, together with the user ratings and the visual features extracted from the movies, is available for download[1].

2. RELATED WORK

Multimedia recommender systems typically exploit *high-level* features in order to generate movie recommendation [5, 15, 6, 16, 7]. This type of features express semantic properties of media content that are obtained from structured sources of meta-information such as databases, lexicons and ontologies, or from less structured data such as reviews, news articles, item descriptions and social tags.

In contrast, in this paper, we propose exploiting *low-level* features to be exploited for recommendation generation. Such features express stylistic properties of the media content and they are extracted directly from multimedia content files [10]. This approach has been already investigated in music recommendation domain [2, 20]. However, it has received marginal attention in movie recommendation domain.

The very few approaches in the video recommendation domain which exploit low-level features only consider scenarios where low-level features are used jointly with high-level features to improve the quality of recommendations. The work in [22] proposes a video recommender system, called *VideoReach*, which incorporate a combination of high-level and low-level video features (such as textual, visual and aural) in order to improve the click-through-rate metric. The work in [24] proposes a multi-task learning algorithm to integrate multiple ranking lists, generated by using different sources of data, including visual content.

This paper addresses a different scenario [12], i.e., when the high-level features are not available (e.g., in the new item scenario). Accordingly, the proposed recommender system can analyze the movies, extract a set of low-level visual features, and use it effectively to generate personalized recommendations.

3. METHOD DESCRIPTION

Video content features can be roughly classified into two hierarchical levels:

High-level (HL): the semantic features that deal with the concepts and events in a movie, e.g. the plot of a movie which consists of a sequence of events.

Low-level (LL): the stylistic features that define the mise-en-scene characteristics of the movie, i.e., the design aspects that characterize aesthetic and style of a movie.

Recommender systems in the movie domain use HL features, usually provided by a group of domain experts or by a large community of users, such as movie genres (structured features, high level). Our focus in this work is mainly on the LL visual features. The influence of these elements in the perception of a movie in the eyes of a viewer has been observed in the works by [4, 13] and were identified in our previous works [10, 11, 9]. The method used to extract low-level visual features and to embed them in movie recommendations is composed of the following steps as shown in Figure 1: (i) Video structure analysis, (ii)Video content analysis, and (iii) Recommendation.

[1]http://recsys.deib.polimi.it/

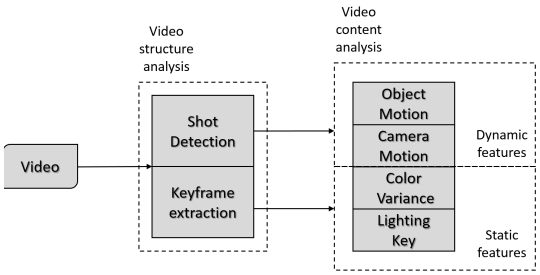

Figure 1: Generic framework of our video analysis system

Video structure analysis aims at segmenting a video into a number of structural elements including shots and key frames. Shots are sequences of consecutive frames captured without interruption by a single camera. From each shot, we extract the middle frame as the representative key frame, inspired by [17]. Two classes of features are extracted in the next stage, i.e., the video content analysis stage: dynamic features capturing the temporal aspects in a video (e.g. object motion and camera motion) are calculated from consecutive frame within each shot and static feature capturing spatial aspect (e.g. color variance and lighting key) are extracted from key frames. The visual feature vector $\mathbf{f_v}$ is composed of the following elements:

$$\mathbf{f_v} = (\overline{L}_{sh}, \mu_{cv}, \sigma^2_{cv}, \mu_{\overline{m}}, \mu_{\sigma^2_m}, \mu_{lk}, n_s) \quad (1)$$

where $\overline{L}_{sh}$ is the average shot length, μ_{cv} and σ^2_{cv} are the mean and the standard deviation of color variance across key frames, $\mu_{\overline{m}}$ and $\mu_{\sigma^2_m}$ are the mean of motion average and the mean of motion standard deviation across all frames, μ_{lk} is the mean lighting key over key frames and n_s is the number of shots.

3.1 Recommendation algorithm

In order to generate recommendations using our low-level visual features, we adopted a complex algorithm called Factorization Machines (FM) [18]. FM is one of the most advanced predictors and it is a combination of well-known Support Vector Machines (SVM) with factorization models. FM is a generic predictor that can work with any feature vector such as our visual features, or genre. It has been already tested and has shown excellent performance in content-based recommendation [14, 8], and high scalability in big datasets [18]. FM computes rating predictions as a weighted combination of the latent factors, low-level visual features, and biases. FM models complicated relationships in the data.

We have used two baselines: genre-based FM, which uses the item feature vector of length 19 (i.e., the number of unique genres), and top-rated non-personalized recommender.

3.2 Normalization

After the extraction of the low-level visual features, they have been normalized adopting 3 types of normalization:

Logarithmic: for every low-level feature (out of 7), the values of that feature is passed through a logarithmic function (natural logarithm). This changed the distributions to be approximately normal, as the original features in the dataset had a distribution similar to log normal distribution.

Quantile: for every low-level feature, the values of that

feature are normalized by applying quantile normalization [3]. This would change the distribution of all the features to be similar.

Log-Quantile: for every low-level feature, the values of that feature are normalized by applying logarithmic normalization (natural logarithm). Then, quantile normalization is applied to make the distribution of all the features to be similar.

Finally, regardless of the normalization type, we scaled the values of all features to the range of 0-1.

4. RESULTS

We have used the latest version of the Movielens dataset which contains 22,884,377 ratings provided by 247,753 users to 34,208 movies (sparsity 99.72%) [1]. For every movie, we queried Youtube and downloaded the trailer, if available. The final dataset contains **13M** ratings provided by **182K users** to **13,373** movies classified along 19 genres: *Action*, *Adventure*, *Animation*, *Children's Comedy*, *Crime*, *Documentary*, *Drama*, *Fantasy*, *Film-Noir*, *Horror*, *Musical*, *Mystery*, *Romance*, *Sci-Fi*, *Thriller*, *War*, *Western*, and *unknown*. Low-level features have been automatically extracted from trailers. We have used trailers and not full videos in order to have a scalable recommender system.

4.1 Experiment 1: Normalization

In order to evaluate the effectiveness of different normalization techniques for low-level features, we have fed the FM model with the low-level visual features. Table 1 presents the results of the evaluation. Different normalizations of the data result in different performances of low level features. Indeed, our observation shows that the best normalization is log-quantile, which by far, outperforms the other two methods, in terms of all evaluation metrics, we considered. The other two methods, have shown similar performance with no significant differences.

These results may point out that the main difference between the best method and the other two is the adoption of both logarithmic and quantile normalizations [3]. Indeed, this may indicate that both logarithmic and quantile normalizations are very necessary to be adopted to obtain the visual low-level feature values that can well represent the movie trailers, and at the same time, distinguish them from each other.

According to these results, hereafter, we only present the results of the best normalization method, i.e., FM visual-low-level recommendation technique based on log-quantile normalization, when performing comparison with the other recommender baselines.

4.2 Experiment 2: Feature Comparison

Table 2 presents the results we have obtained from the conducted experiments. As it can be seen, in terms of Precision, by far, the best technique is our proposed visual low-level feature based FM. Our technique obtained precision scores 0.0367, 0.0343, and 0.0286, while genre-based technique obtained scores of 0.0041, 0.0038 and 0.0040, for different recommendation size K at 5, 10, and 30. This result is promising since it shows that our technique based on automatic extraction of low-level visual features can achieve precision scores much better than genre-based recommendation.

Similar result has been observed w.r.t. recall metric. In terms of recall, our proposed technique, again similarly, have achieved the best result. While recall scores of our technique are 0.0272, 0.0488, 0.1176, genre-based FM obtained 0.0025, 0.0049, and 0.0170 for K at 5, 10, and 30. As expected, the non-personalized top-rated recommendation technique is the worst technique among all in terms of both precision and recall metrics.

We have also computed the F1 metric. Comparing the results, our proposed technique outperforms all the other technique in terms of F1. It achieves 0.0312, 0.0403, and 0.0461 scores and genre-based FM achieves 0.0031, 0.0043, and 0.0068 for K at 5, 10, and 30, which is substantially greater than the genre-based technique. Again top-rated technique achieves the worst result in terms of F1.

Comparing all these promising results, it is clear that our proposed technique, i.e., recommendation based on FM algorithm incorporating automatically extracted low-level visual features performs almost 10 times better scores than the recommendation based on rich source of expert-annotated genre labels, in terms of precision, recall , and F1 metrics.

5. CONCLUSION AND FUTURE WORK

This work presents a novel approach in the domain of content-based movie recommendations. The technique is based on the analysis of movie content and extraction of low-level visual features, fed to the Factorization Machine algorithm as side information, in order to generate personalized recommendations for users. This approach makes it possible to recommend items to users without relying on any high-level semantic features (e.g., genre) that are expensive to obtain, as they require expert level knowledge, and shall be missing (e.g., in new item scenario).

The results of our evaluation show that recommendations based on low-level visual features achieves almost 10 times better accuracy in comparison to the recommendations based on traditional set of high-level semantic features (i.e., genre).

For future work, we consider the design and development of an online web application in order to conduct online studies with real user. The goal is to evaluate the effectiveness of recommendations based on low-level visual features not only in terms of relevance, but also in terms of novelty, diversity and serendipity. Moreover, we will extend the range of low-level features extracted, and also, include audio features. Finally, we will extend the evaluation to user-generated videos.

6. ACKNOWLEDGMENTS

This work is supported by Telecom Italia S.p.A., Open Innovation Department, Joint Open Lab S-Cube, Milan.

7. REFERENCES

[1] Datasets | grouplens. http://grouplens.org/datasets/. Accessed: 2015-05-01.

[2] D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *Multimedia, IEEE Transactions on*, 13(4):687–701, 2011.

[3] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

Table 1: Performance comparison of different normalizations of visual low-level features

Recommender	Norm	Precision			Recall			F1		
		@5	@10	@30	@5	@10	@30	@5	@10	@30
Visual-ll feature FM	**log-quantile**	**0.0367**	**0.0343**	**0.0286**	**0.0272**	**0.0488**	**0.1176**	**0.0312**	**0.0403**	**0.0461**
	quantile	0.0083	0.0081	0.0075	0.0052	0.0104	0.0306	0.0064	0.0091	0.0121
	log	0.0084	0.0084	0.0083	0.0051	0.0101	0.0308	0.0063	0.0092	0.0131

Table 2: Performance comparison of various recommendation techniques

Recommender	Precision			Recall			F1		
	@5	@10	@30	@5	@10	@30	@5	@10	@30
Visual-ll feature FM	**0.0367**	**0.0343**	**0.0286**	**0.0272**	**0.0488**	**0.1176**	**0.0312**	**0.0403**	**0.0461**
Genre-based FM	0.0041	0.0038	0.0040	0.0025	0.0049	0.0170	0.0031	0.0043	0.0068
Top rated	1.390e-05	1.042e-05	1.040e-05	1.922e-06	3.087e-06	1.111e-05	3.377e-06	4.764e-06	1.074e-05

[4] D. Bordwell, K. Thompson, and J. Ashton. *Film art: An introduction*, volume 7. McGraw-Hill New York, 1997.

[5] I. Cantador, M. Szomszor, H. Alani, M. Fernández, and P. Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. 2008.

[6] M. De Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. Semantics-aware content-based recommender systems. In *Recommender Systems Handbook*, pages 119–159. Springer, 2015.

[7] M. De Gemmis, P. Lops, and G. Semeraro. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, 17(3):217–255, July 2007.

[8] Y. Deldjoo, M. Elahi, and P. Cremonesi. How to combine visual features with tags to improve the movie recommendation accuracy. In *International Conference on Electronic Commerce and Web Technologies*. Springer International Publishing, 2016.

[9] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, and P. Piazzolla. Recommending movies based on mise-en-scene design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1540–1547. ACM, 2016.

[10] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, pages 1–15, 2016.

[11] Y. Deldjoo, M. Elahi, M. Quadrana, and P. Cremonesi. Toward building a content-based video recommendation system based on low-level features. In *E-Commerce and Web Technologies*. Springer, 2015.

[12] M. Elahi, F. Ricci, and N. Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 2016.

[13] J. Gibbs. *Mise-en-scène: Film style and interpretation*, volume 10. Wallflower Press, 2002.

[14] L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 557–566, New York, NY, USA, 2013. ACM.

[15] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. In *User Modeling, Adaptation, and Personalization*, pages 188–199. Springer, 2012.

[16] M. Nasery, M. Elahi, and P. Cremonesi. Polimovie: a feature-based dataset for recommender systems. In *ACM RecSys Workshop on Crowdsourcing and Human Computation for Recommender Systems (CrawdRec)*, volume 3, pages 25–30. ACM, 2015.

[17] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):52–64, 2005.

[18] S. Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[19] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan. Active learning in recommender systems. In *Recommender Systems Handbook - chapter 24: Recommending Active Learning*, pages 809–846. Springer US, 2015.

[20] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2010.

[21] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1):3, 2014.

[22] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80. ACM, 2007.

[23] H. Zettl. Essentials of applied media aesthetics. In C. Dorai and S. Venkatesh, editors, *Media Computing*, volume 4 of *The Springer International Series in Video Computing*, pages 11–38. Springer US, 2002.

[24] X. Zhao, G. Li, M. Wang, J. Yuan, Z.-J. Zha, Z. Li, and T.-S. Chua. Integrating rich information for video recommendation with multi-task rank aggregation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1521–1524. ACM, 2011.