Recommending Videos in Cold Start With Automatic Visual Tags

Mehdi Elahi
mehdi.elahi@uib.no
University of Bergen
Norway

Farshad Bakhshandegan Moghaddam
farshad.moghaddam@uni-bonn.de
University of Bonn
Germany

Reza Hosseini
seyed-reza.hosseini@vaillant-group.com
Vaillant Group
Germany

Mohammad Hossein Rimaz
hosseinrimaz@gmail.com
University of Passau
Germany

Nabil El Ioini
nelioini@unibz.it
Free University of Bozen - Bolzano
Italy

Marko Tkalčič
marko.tkalcic@gmail.com
University of Primorska
Slovenia

Christoph Trattner
Christoph.Trattner@uib.no
University of Bergen
Norway

Tammam Tillo
tammam@iiitd.ac.in
Indraprastha Institute of Information Technology - Delhi
India

ABSTRACT

This paper addresses the so-called *New Item* problem in video Recommender Systems, as part of *Cold Start*. New item problem occurs when a new item is added to the system catalog, and the recommender system has no or little data describing that item. This could cause the system to fail to meaningfully recommend the new item to the users. We propose a novel technique that can generate cold start recommendation by utilizing automatic *visual* tags, i.e., tags that are automatically annotated by deeply analyzing the content of the videos and detecting faces, objects, and even celebrities within the videos. The automatic visual tags do not need any human involvement and have been shown to be very effective in representing the video content. In order to evaluate our proposed technique, we have performed a set of experiments using a large dataset of videos. The results have shown that the automatically extracted visual tags can be incorporated into the cold start recommendation process and achieve superior results compared to the recommendation based on human-annotated tags.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

Recommender Systems, Visual Tags, Visual Features, Cold Start

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Mehdi Elahi, Farshad Bakhshandegan Moghaddam, Reza Hosseini, Mohammad Hossein Rimaz, Nabil El Ioini, Marko Tkalčič, Christoph Trattner, and Tammam Tillo. 2021. Recommending Videos in Cold Start With Automatic Visual Tags. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A major challenge in Recommender Systems is known as the *New Item* problem. This problem is part of a bigger challenge called *Cold Start* problem, and it occurs when a new item is added to the item catalog and no rating has been provided by the users for that new item [19, 30, 40, 52]. Content-Based Filtering (CBF) is a recommendation technique that can alleviate the cold start problem by using the item metadata (e.g., tags) to find items that have (content-wise) similarities and to recommend to a target user the items that are similar to those items liked by the user in the past [15, 33, 43, 50]. This technique can be used in a variety of application domains, including the video domain, where the video content is represented by *high-level* features (e.g., tags added to videos) and *low-level* features (e.g., colorfulness in videos) [16]. The former type of content features represents the *semantics* illustrated by the concepts and events happening within a video (e.g., Titanic 1997 annotated with *#LeonardoDiCaprio* tag) [23]. The latter type of features, on the other hand, represents the *stylistic* aspects of videos defined by the aesthetic characteristics of the videos (e.g., Alice in Wonderland 2010 having a high value of *colorfulness*). Content-based video recommendation typically focuses on exploiting *high-level* features, which can be either *manual* (annotated by human) or *automatic* (annotated by algorithms). While manual features can be informative descriptors of an item, they are typically either unavailable or expensive to collect. As an example, in the world's biggest online video community (*YouTube*) the videos are often uploaded with no or very poor metadata [10].

In the cold start scenario, where a new item is added to the system catalog and no user has yet rated the new item or annotated

it with any metadata, the recommender system may be unable to generate a relevant recommendation of the new item. An example can be a video uploaded to a video-sharing platform and none of the users has yet rated or tagged that video. In many cases, even the video maker herself forgets to include a meaningful description when uploading the video. In any of the above cold start situations, even sophisticated recommendation algorithms may fail to make relevant recommendations.

This paper proposes a novel recommendation approach based on automatic *visual* tags. Such features are automatically identified and added to the video items using *Deep Learning* models [5, 14, 57, 58]. Examples of visual tags are a set of tags automatically added to a video, representing objects, faces, and celebrities within that video. These visual tags can describe the *high-level* and semantic content of the video file (e.g., *#AngelinaJolie in #Airplane)*, in contrast to visual features describing *low-level* and stylistic content (e.g., *colorfulness* and *brightness*). Our proposed visual tags are then used to generate content-based video recommendation for users and compared against (manual) tags that need human annotation, and not necessarily always available.

We have performed a number of experiments using a large dataset of 7,689 movie trailers in order to evaluate the quality of recommendation based on (automatic) visual tags. We used movie trailers since prior works have shown high visual similarity between the trailers and their corresponding full-length movies [11]. In these experiments, we have compared the recommendation based on "combined" high-level visual tags with the recommendation based on each "individual" type of them, i.e., *Celebrity* tags, *Facial* tags, and *Object* tags. We compared the quality of these (automatic) tags with recommendations based on (manual) tags as well as recommendations based on low-level features. The results have shown the effectiveness of our proposed visual tags, in comparison to the recommendation based on (manual) tags and low-level features. To the best of our knowledge, this is the first attempt for generating (high-level) visual tags and comparing it against alternative (low-level) visual features considering the (extreme) cold start situations when no other type of content data exists for the videos.

It is worth noting that we primarily focused on recommendation based on tags as prior studies have shown the superior performance of tags in comparison to other types of content features (e.g., genre) [13]. Furthermore, using visual tags enables the recommender systems to include explanation when presenting recommended videos to their users. Explanation may enhance transparency of the system and result in higher user satisfaction [54]. This is not very feasible with the pure *low-level* visual features.

The main contributions of this paper is listed in the following:

- we have extracted a large dataset of (automatic) visual tags from 7,689 movie trailers, using Deep Learning models, capable of annotating movies with a wide range of automatic tags including celebrity tags (e.g., *#TomHanks & #BradPitt*), object tags (e.g., *#sky & #children*), and face tags (e.g., *#happy & #withGlass*); our dataset is public and freely available on Figshare[1];
- we have addressed the cold start problem by proposing a novel set of content features, extracted automatically, with no need for any human involvement and used for recommending new items with no rating and no tags;
- we have evaluated the proposed content-based recommendation approach using a large dataset of thousands of movie trailers and compared our results with different baselines, including recommendation based on *low-level* visual features (e.g., colorfulness, sharpness, and naturalness in movies);
- our results have shown the superiority of recommendation based on our *high-level* visual tags, used all together or individually (e.g., using only celebrity tags), in comparison to the manual tags and *low-level* visual features.

[1] https://doi.org/10.6084/m9.figshare.14528727

2 BACKGROUND

In this section, we will briefly review two related research areas, i.e., (a) tag-based recommender systems and (b) visually-aware recommender systems. Several prior works have incorporated (human-annotated) tags into the recommendation process[1, 2, 6, 12, 20, 21, 36]. An example can be [29] where integrating tag-based similarities within a Collaborative Filtering system has yielded an improvement in the recommendation. Another example can be [3] where user tags and item descriptions have been incorporated in the recommendation process. Another example can be [17] where the authors have proposed a modified version of SVD++ Matrix Factorization model [28] by replacing the usage of implicit feedback with tagging information. This has resulted in a substantial improvement of the recommendation performance. In [35] and [18] the matrix factorization model has been extended with incorporation of latent factors associated to the item features.

Recent works have proposed using different forms of visual features for recommendation that can be grouped into two classes, i.e., (i) *low-level* features (typically based on hand-crafted approaches) and (ii) *high-level* features (typically based on deep learning approaches) [24, 25, 27, 31]. The usage of the low-level visual features has drawn minor attention in recommender systems (e.g., in [11, 16, 37, 38, 45]). This is while this has been extensively investigated in the other fields such as computer vision [41, 51]. [7, 26] provided comprehensive surveys on the state-of-the-art techniques related to the video content analysis and discussed several low-level features (e.g. visual, textual, or auditory) that can be used for various applications, including classification or recommendation. An example of the works using such features is [44] where a framework for movie genre classification based only on visual features has been proposed. [53] proposed a deep learning approach to automatically detect the director of a movie based on low-level visual features.

It is worth noting that, while hand-crafted features [27] may still offer promising performance, recently, deep learning-based approaches have achieved a superior accuracy in comparison to them [8]. Convolutional Neural Networks (CNN) is an example of effective deep learning approaches that can build a informative representation of items [32].

This work differs from the prior works as it proposes high-level (automatic) *visual* tags instead of pure low-level visual features. One of main differences is that we have focused on the task of video recommendation while many prior works focused on video annotation or labeling (e.g., [42, 49]). Another difference is that

our work can be used when generating explanations for the recommendation due to the high-level nature of the proposed visual tags which makes them to be human-understandable compared to the low-level features. Finally, we have used a large-scale dataset for our evaluation with thousands of items compared to some of the prior works which used small-scale datasets (e.g., [11] considering only few hundreds of items).

3 PROPOSED METHOD

This section explains how our two datasets are generated, one contains the *low-level features* (i.e. colorfulness, sharpness, saturation, etc.) and another contains *high-level features* (i.e. celebrity tags, object & label tags, and face tags). First of all, we used a large dataset of movie trailers, obtained through querying YouTube based on the movie titles in the MovieLens dataset [22]. Prior works have shown a high similarity of visual features extracted from movie trailers and their respective full-length movies [11]. After an initial prepossessing, we have extracted visual features from 7,689 movie trailers, conducting the following steps: *Movie Segmentation*, *Feature Extraction*, and, *Feature Aggregation*.

3.1 Movie Segmentation.

In order to segment movies into shots, i.e., sequences of consecutive frames recorded without camera interference, we used a method based on *Color Histogram Distance* [4]. This is due to the fact that the transition between two shots of the video is typically very abrupt, and hence, the color histogram differences among the movie frames can be an indicative of it. Finally, for every shot, the middle frame is selected as the key-frame.

3.2 Feature Extraction

3.2.1 ***Low-Level Features:*** [2] We have extracted a set of low-level visual features capable of effectively capturing the *attractiveness* of each key frame within the movies. A prior work [47] showed that these features can be well indicative of how attractive the Flicker images are. Table 1 (top half) summarizes the full set of extracted features.

3.2.2 ***High-Level Features:*** we have extracted another dataset containing a novel set of high-level features in the form of visual tags (labels). The main advantage of these novel features over the low-level features is that high-level features are human-understandable and hence sound meaningful to the users. This enables them to be exploited for various purposes, e.g., generating explanation of recommendation for users or automatically creating a brief summary of the movies. It is worth noting that, to the best of our knowledge, this is the first time that a large movie dataset with **(i) a collection of powerful content descriptors consisted of both high-level visual tags & low-level visual features**, being (ii) directly linked to millions of user ratings and tags is published and accessible for the community. For creating this dataset, we initially considered exploiting the Deep Learning approaches and frameworks such as *ImageAI*[3], *OpenCV*[4], and *MTCNN*[5]. However, we have encountered a number of challenges needed to be tackled. The main challenge concerned the low quality of the movie trailers (and hence their corresponding key frames) we obtained for some of the old movies. As a consequence, this has yielded in lower quality of the extracted visual tags. Hence, we checked alternative enterprise services and found them to be more robust compared to the above-mentioned open-source approaches. Hence, we decided to opt for a paid cloud-based service offered by Amazon Web Services (AWS). The service is called *Rekognition* [6] which is a Software as a Service (SaaS) computer vision platform capable of extracting a large number of visual tags, as well as their corresponding confidence scores in the range of 0%-100%. Table 1 (bottom half) shows the extracted high-level visual tags for each movie. **Celebrity Tags**: Rekognition can recognize thousands of celebrity individuals who are famous, noteworthy, or prominent in their field. **Object Tags (Labels)**: Rekognition can detect a wide range of labels within the movies such as vehicles, pets, natural objects, office equipments, buildings, and etc. **Face Tags (Facial attributes)**: Rekognition is able to locate faces within images and analyze face attributes, such as whether or not the face is smiling or the eyes are open. It can also detect emotions, namely, 'happy', 'sad', 'angry', 'confused', 'disgusted', 'surprised', 'calm', 'fear', and 'unknown'.

3.3 Feature Aggregation

To form the feature vector description of a movie, we used a combination of term frequency–inverse document frequency (*tf-idf*) method and *Word2Vec* vectors [39] trained on GoogleNews[7] as the following. First, we collected all the celebrities detected within the set of all frames of each movie. Considering each movie as a document and each label as a word, we calculated the tf-idf scores of each word. In addition to that, we computed the vector representation of each word using the Word2Vec network. This is a real-value vector of length 300. In order to make a single vector for each movie, we calculated the weighted average of vector-representations of all the celebrity tags appeared in the movie with tf-idf values as their weight.

3.4 Recommendation algorithm

We adopted a classical "K-Nearest Neighbor" content-based algorithm. Given a set of users $u \in U$ and a catalog of items $i \in I$, a set of preference scores r_{ui} given by user u to item i has been collected. Each item $i \in I$ is associated to its feature vector $\boldsymbol{f_i}$. For each couple of items i and j, the similarity score s_{ij} is computed using *cosine similarity* and utilized for rating prediction:

$$s_{ij} = \frac{\boldsymbol{f_i}^T \boldsymbol{f_j}}{\boldsymbol{f_i}\boldsymbol{f_j}} \qquad \hat{r}_{ui} = \frac{\sum_{j \in NN_i, r_{uj}>0} r_{uj} s_{ij}}{\sum_{j \in NN_i, r_{uj}>0} s_{ij}} \tag{1}$$

where NN_i is the set of nearest neighbors for each item i.

[2] https://www.researchgate.net/publication/333579748_MA14KD_AGGREGATED_Dataset_Description_Visual_Attraction_of_Movie_Trailers

[3] https://github.com/OlafenwaMoses/ImageAI
[4] https://opencv.org/
[5] https://github.com/ipazc/mtcnn
[6] https://aws.amazon.com/rekognition/
[7] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Table 1: Characteristics of two datasets extracted from movie trailers

Dataset	Feature	Description	Details
Low-Level Features (Colorfulness, (Brightness, etc.)	Sharpness	level of details within a frame	#Features=10
	Sharpness Variation	standard deviation of all pixel sharpness values	
	Contrast	relative difference in brightness/color of features	
	RGB Contrast	contract which is extended to RGB color space	
	Saturation	colorfulness relative to brightness	
	Saturation variation	standard deviation of all pixel saturation values	
	Brightness	average brightness of a frame	
	Colorfulness	individual color distance of pixels in a frame	
	Entropy	amount of information in a video frame	
	Naturalness	difference between a frame & human perception	
High-Level (Celebrity, Object, Face)	celebrity_name	name of detected celebrity	#Celebrities=29,132
	celebrity_url	URL of imdb page for celebrity (can be empty)	
	match_confidence	confidence rate [50%,100%]	
	label_confidence	confidence rate [0%,100%]	#Labels=2,636
	face_conf	confidence rate [0%,100%]	#Features=28
	age_range	age range of detected face	
	emotion	level of confidence in determination	
	gender_info	gender value and confidence level of gender detection	
	eyeglasses/sunglasses	true, false and confidence level of a eye glass/sunglasses detection	
	eyesopen_info	true, false and confidence level of an eye open detection	
	smile_info	true, false and confidence level of a smile detection	
	mouthopen_info	true, false and confidence level of a mouth open detection	
	mustache/beard	true, false and confidence level of mustache/beard detection	

4 EXPERIMENTAL RESULTS

4.1 Methodology

For evaluation, we followed a methodology similar to the one proposed by [9]. We used a large rating dataset, i.e., MovieLens with 25M ratings, and filtered out users who have rated at least 10 relevant items (i.e., items with ratings equal or higher than 4). This ensured us that each user has a minimum number of favorite items. For each selected user, we chose 2 items with rating equal or higher than 4 (forming a favorite set of items). Then we randomly added 500 items not rated by the user to this set. After that we predicted the ratings for all the 502 movies using the recommender system and ordered them according to the predicted ratings. For each $1 \leq N \leq 502$, the number of hits is the number of favorite movies appeared in top N movies (e.g. 0, 1 or 2). Assume T is the total number of favorite items in the test set for all selected users ($T = 8000$ in our case), then:

$$recall@N = \frac{\#hits}{T} \qquad precision@N = \frac{\#hits}{N \cdot T} = \frac{recall@N}{N} \quad (2)$$

4.2 Visualizing Automatic Tags

For the aim of visualization of the data, we used a powerful dimensionality reduction method called *T-distributed Stochastic Neighbor Embedding method (t-SNE)* [34]. The result has been plotted in Figure 1. Please note that, every point in this figure represents a tag and the distances are indicative of the visual similarities. Hence, tags could be positioned close to or far from each other, depending on their visual similarities. As it is seen in the figure, although the distances are computed based on visual similarities (which is not necessarily translated to pure tag semantics), however, the tags that are located close by are semantically related. For example, as seen in the figure, the following tags located in the bottom right side of the figure are semantically related: *dark, detective, horror, life & death, murder* and *mystery*.

4.3 Recommendation in Cold Start

We exploited different forms of low-level visual features and high-level visual tags (see Table 1), extracted automatically from video to build a content-based recommender system. We evaluated the system considering the *new item* cold start scenario. It is worth noting that, in the severe cold start scenario, a video item may have neither any rating nor any *manual* tag. In such a case, the system can only rely on our proposed (automatic) visual tags, as they require no human-annotation. In the moderate scenario of cold start, a limited number of users may have added few manual tags, and the recommender system can generate personalized recommendation based on them.

We have evaluated the performance of our proposed recommender system using (automatic) visual tags in terms of precision@N, and recall@N [48]. Although we have also computed F1@N scores, due to the space limit we have not reported these results. Moreover, as the main baseline, we have considered the recommendation based on tags since prior studies have proven the superior performance of tags in comparison to the other types of content features (e.g., genre) [13].

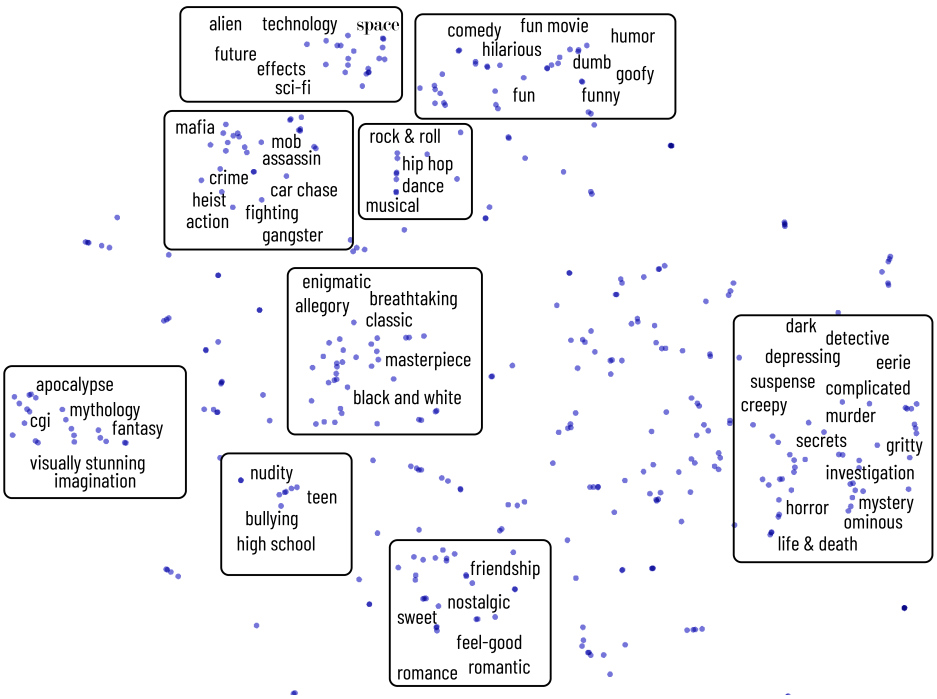

Figure 1: Analyzing user-annotated tags, based on visual features within the videos, by applying *t-SNE* technique.

Figure 2 presents the results in terms of precision@N (left sub-figure) and recall@N (right sub-figure). As it can be seen, by far the best result has been achieved by recommendation based on (automatic) celebrity tags, annotated based on Deep Learning model, and for all range of recommendation size (1<N<20). The precision values started at 0.013 for precision@1 and reached the value of 0.004 for precision@20. The second best performance has been observed for recommendation based on the automatic visual tags (i.e., combination of Celebrity+Facial+Object tags). Recommendation based on manual tags (human-annotated) has shown to have the third best performance among all features up to the N=5. However, when N got larger than 5, *low-level* visual features (e.g., colorfulness, sharpness, naturalness, etc.) has outperformed the manual tags.

Similar results have been observed for the recall metric. As it can be seen, again, the recommendation based on (automatic) celebrity tags has expressed substantially better performance by achieving the highest recall values for all different recommendation sizes (N). The recall values for this method has begun with 0.013 for recall@1 and reached 0.084 for recall@20. The next best performance is observed for automatic visual tags. Recommendation based on visual tags has achieved 0.009 for recall@1 and 0.048 for recall@20. Recommendation based on manual tags has not not been very different from visual tags where the values are 0.007 for recall@1 and 0.037 for recall@20. For both precision@N and recall@N the worse performance has been achieved by (automatic) object and (automatic) facial tags. This can be due to our particular aggregation methodology and can be substantially improved by using a novel feature fusion technique. Despite the observed poor performance of these type of automatic features, however, these features can still serve as a potential solution for cold start scenario where no tag and no rating has been available for a new item.

5 CONCLUSIONS & FUTURE WORK

In this paper, we address the so-called cold start challenge in recommender systems and propose a technique to generate recommendation based on *visual* tags. These are novel features that describe the video content and can be automatically annotated and used when a new item has not received any rating or any user tag. In such a severe case, any form of complicated recommender algorithm may fail to generate relevant recommendations. We have also performed experiments, assuming that the users have *manually* annotated a number of tags. The results revealed a superior quality of recommendations based on visual tags compared to the manual tags. These results are promising as they demonstrate the potential power of visual tags in dealing with severe cases of cold start problem.

Our future work plan includes implementing a new component that can analyze facial expressions of users and collect user preferences from such novel form of data [55, 56]. In addition to that, we plan to extend our feature set by including audio features collected in a recent work [46]. This will enable our proposed technique to generate recommendations based on a novel set of audio-visual features.

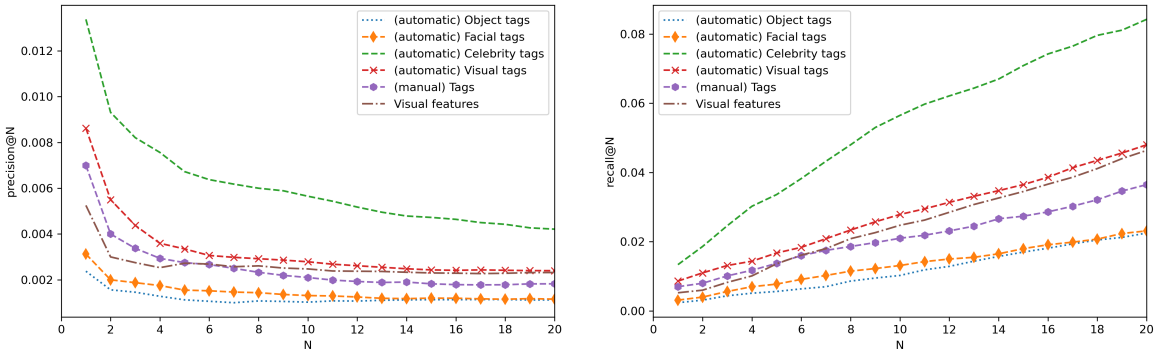

Figure 2: Comparing recommendation based on different features, in terms of (left) precision@N and (right) recall@N

ACKNOWLEDGMENTS

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through The Centres for Research-based Innovation scheme, project number 309339.

REFERENCES

[1] Benjamin Adrian, Leo Sauermann, and Thomas Roth-Berghofer. 2007. Contag: A semantic tag recommendation system. *Proceedings of I-Semantics* 7 (2007), 297–304.

[2] Syed M Ali, Gopal K Nayak, Rakesh K Lenka, and Rabindra K Barik. 2018. Movie recommendation system using genome tags and content-based filtering. In *Advances in Data and Information Sciences*. Springer, 85–94.

[3] Fahad Anwar, Naima Iltaf, Hammad Afzal, and Haider Abbas. 2019. A Deep Learning Framework to Predict Rating for Cold Start Item Using Item Metadata. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. IEEE, 313–319.

[4] Edoardo Ardizzone, Marco La Cascia, and Davide Molinelli. 1996. Motion and color-based video indexing and retrieval. In *Proceedings of 13th International Conference on Pattern Recognition*, Vol. 3. IEEE, 135–139.

[5] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. 2016. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security* 11, 9 (2016), 1903–1913.

[6] Toine Bogers. 2018. Tag-based recommendation. In *Social Information Access*. Springer, 441–479.

[7] D. Brezeale and D. J. Cook. 2008. Automatic Video Classification: A Survey of the Literature. *Trans. Sys. Man Cyber Part C* 38, 3 (May 2008), 416–430. https://doi.org/10.1109/TSMCC.2008.919173

[8] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. 2015. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5315–5324.

[9] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. 39–46.

[10] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. 293–296.

[11] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-Based Video Recommendation System Based on Stylistic Visual Features. *Journal on Data Semantics* (2016), 1–15. https://doi.org/10.1007/s13740-016-0060-9

[12] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Farshad Bakhshandegan Moghaddam, and Andrea Luigi Edoardo Caielli. 2016. How to combine visual features with tags to improve movie recommendation accuracy?. In *International conference on electronic commerce and web technologies*. Springer, 34–45.

[13] 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* 7, 4 (2018), 207–219. https://doi.org/10.1007/s13735-018-0155-1

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[15] Mehdi Elahi, Matthias Braunhofer, Tural Gurbanov, and Francesco Ricci. 2018. User Preference Elicitation, Rating Sparsity and Cold Start.

[16] 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 Conference on Recommender Systems*. 326–330.

[17] Manuel Enrich, Matthias Braunhofer, and Francesco Ricci. 2013. Cold-Start Management with Cross-Domain Collaborative Filtering and Tags. In *Proceedings of the 13th International Conference on E-Commerce and Web Technologies*. Springer, 101–112. https://doi.org/10.1007/978-3-642-39878-0_10

[18] Ignacio Fernández-Tobías and Iván Cantador. 2014. Exploiting Social Tags in Matrix Factorization Models for Cross-domain Collaborative Filtering. In *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems, Foster City, California, USA*. 34–41.

[19] Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. 2010. Learning attribute-to-feature mappings for cold-start recommendations. In *2010 IEEE International Conference on Data Mining*. IEEE, 176–185.

[20] Mouzhi Ge, Mehdi Elahi, Ignacio Fernaández-Tobías, Francesco Ricci, and David Massimo. 2015. Using tags and latent factors in a food recommender system. In *Proceedings of the 5th International Conference on Digital Health 2015*. 105–112.

[21] Fatih Gedikli and Dietmar Jannach. 2013. Improving recommendation accuracy based on item-specific tag preferences. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 1 (2013), 1–19.

[22] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[23] Naieme Hazrati and Mehdi Elahi. 2020. Addressing the New Item problem in video recommender systems by incorporation of visual features with restricted Boltzmann machines. *Expert Systems* (2020), e12645.

[24] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[25] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[26] Weiming Hu, Nianhua Xie, Li, Xianglin Zeng, and Stephen Maybank. 2011. A Survey on Visual Content-Based Video Indexing and Retrieval. *Trans. Sys. Man Cyber Part C* 41, 6 (Nov. 2011), 797–819. https://doi.org/10.1109/TSMCC.2011.2109710

[27] Shatha Jaradat. 2017. Deep cross-domain fashion recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 407–410.

[28] Yehuda Koren and Robert Bell. 2011. *Advances in Collaborative Filtering*. Springer US, Boston, MA, 145–186. https://doi.org/10.1007/978-0-387-85820-3_5

[29] Huizhi Liang, Yue Xu, Yuefeng Li, and Richi Nayak. 2009. Tag Based Collaborative Filtering for Recommender Systems. In *Rough Sets and Knowledge Technology, 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14-16, 2009. Proceedings*. 666–673. https://doi.org/10.1007/978-3-642-02962-2_84

[30] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert Systems with Applications* 41, 4 (2014), 2065–2073.

[31] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3330–3337.

[32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[33] Pasquale Lops, Dietmar Jannach, Cataldo Musto, Toine Bogers, and Marijn Koolen. 2019. Trends in content-based recommendation. *User Modeling and User-Adapted Interaction* 29, 2 (2019), 239–249.

[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[35] Marcelo Garcia Manzato. 2013. GSVD++: Supporting Implicit Feedback on Recommender Systems with Metadata Awareness *(SAC '13)*. Association for Computing Machinery, New York, NY, USA, 908–913. https://doi.org/10.1145/2480362.2480536

[36] David Massimo, Mehdi Elahi, Mouzhi Ge, and Francesco Ricci. 2017. Item contents good, user tags better: Empirical evaluation of a food recommender system. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 373–374.

[37] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2019. Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction* 29, 2 (2019), 251–290.

[38] Pablo Messina, Vicente Dominquez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2018. Exploring Content-based Artwork Recommendation with Metadata and Visual Features. *User Modeling and User-Adapted Interaction (UMUAI)* 29, 2 (July 2018), 251–290. https://doi.org/10.1007/s11257-018-9206-9

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[40] Farshad Bakhshandegan Moghaddam and Mehdi Elahi. 2019. Cold start solutions for recommendation systems. *Big Data Recommender Systems, Recent Trends and Advances. IET* (2019).

[41] H. R. Naphide and Thomas Huang. 2001. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia* 3, 1 (March 2001), 141–151. https://doi.org/10.1109/6046.909601

[42] Abhishek A Patwardhan, Santanu Das, Sakshi Varshney, Maunendra Sankar Desarkar, and Debi Prosad Dogra. 2019. ViTag: Automatic video tagging using segmentation and conceptual inference. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 271–276.

[43] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.

[44] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. 2005. On the Use of Computable Features for Film Classification. *IEEE Trans. Cir. and Sys. for Video Technol.* 15, 1 (Jan. 2005), 52–64. https://doi.org/10.1109/TCSVT.2004.839993

[45] Mohammad Hossein Rimaz, Mehdi Elahi, Farshad Bakhshandegan Moghadam, Christoph Trattner, Reza Hosseini, and Marko Tkalčič. 2019. Exploring the Power of Visual Features for the Recommendation of Movies *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 303–308. https://doi.org/10.1145/3320435.3320470

[46] Mohammad H Rimaz, Reza Hosseini, Mehdi Elahi, and Farshad Bakhshandegan Moghaddam. [n.d.]. AudioLens: Audio-Aware Video Recommendation for Mitigating New Item Problem. ([n. d.]).

[47] Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and Classifying Attractiveness of Photos in Folksonomies *(WWW '09)*. Association for Computing Machinery, New York, NY, USA, 771–780. https://doi.org/10.1145/1526709.1526813

[48] 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* 7, 2 (2018), 95–116.

[49] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 279–287.

[50] Jiangbo Shu, Xiaoxuan Shen, Hai Liu, Baolin Yi, and Zhaoli Zhang. 2018. A content-based recommendation algorithm for learning resources. *Multimedia Systems* 24, 2 (2018), 163–173.

[51] Cees G.M. Snoek and Marcel Worring. 2005. Multimodal Video Indexing: A Review of the State-of-the-art. *Multimedia Tools and Applications* 25, 1 (01 Jan 2005), 5–35. https://doi.org/10.1023/B:MTAP.0000046380.27575.a5

[52] Hridya Sobhanam and AK Mariappan. 2013. Addressing cold start problem in recommender systems using association rules and clustering technique. In *2013 International Conference on Computer Communication and Informatics*. IEEE, 1–5.

[53] Michele Svanera, Mattia Savardi, Alberto Signoroni, András Bálint Kovács, and Sergio Benini. 2018. Who is the director of this movie? Automatic style recognition based on shot features. *CoRR* abs/1807.09560 (2018). arXiv:1807.09560 http://arxiv.org/abs/1807.09560

[54] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

[55] Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2017. A Research Tool for User Preferences Elicitation with Facial Expressions *(RecSys '17)*. ACM, New York, NY, USA, 353–354. https://doi.org/10.1145/3109859.3109978

[56] Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2019. Prediction of Music Pairwise Preferences from Facial Expressions *(IUI '19)*. ACM, New York, NY, USA, 150–159. https://doi.org/10.1145/3301275.3302266

[57] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3733–3742.

[58] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.