Visually-Aware Video Recommendation in the Cold Start

Conference Paper · July 2020

DOI: 10.1145/3372923.3404778

5 authors, including:

Mehdi Elahi
University of Bergen
97 PUBLICATIONS **1,853** CITATIONS
SEE PROFILE

Reza Hosseini
11 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Mohammad Hossein Rimaz
Universität Passau
6 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Farshad Bakhshandegan Moghaddam
University of Bonn
29 PUBLICATIONS **139** CITATIONS
SEE PROFILE

Visually-Aware Video Recommendation in the Cold Start

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

Reza Hosseini
Vaillant Group Business Services
Remscheid, Germany
seyed-reza.hosseini@vaillant-group.com

Mohammad H. Rimaz
University of Passau
Passau, Germany
mhrimaz@email.kntu.ac.ir

Farshad B. Moghaddam
Karlsruhe Institute of Technology
Karlsruhe, Germany
farshad.moghaddam@student.kit.edu

Christoph Trattner
University of Bergen
Bergen, Norway
christoph.trattner@uib.no

ABSTRACT

Recommender Systems (RSs) have become essential tools in any video-sharing platforms (such as YouTube) by generating video suggestions for users. Although, RSs have been effective, however, they suffer from the so-called *New Item* problem. New item problem, as part of *Cold Start* problem, happens when a new item is added to the system catalogue and the RS has no or little data available for that new item. In such a case, the system may fail to meaningfully recommend the new item to any user.

In this paper, we propose a novel recommendation technique based on *visual* tags, i.e., tags that are automatically annotated to videos based on visual description of videos. Such visual tags can be used in an *extreme* cold start situation, where neither any rating, nor any tag is available for the new video item. The visual tags could also be used in the *moderate* cold start situation when the new video item has been annotated with few tags. This type of content features can be extracted automatically without any human involvement and have been shown to be very effective in representing the video content.

We have used a large dataset of videos and shown that 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.

1 INTRODUCTION

One of the main challenges in Recommender Systems (RS) is the *New Item* problem. This problem which is part of a bigger challenge called *Cold Start* problem happens when a new item is added to the item catalogue and no rating has been provided by the users for that item [13, 32]. In such a case, the RS may fail to effectively recommend that item to users.

One of the recommendation techniques that can remedy the cold start problem is Content-Based Filtering (CBF) which can exploit content data (e.g., item tag) in order to compute similarity among items and generate relevant recommendation based on content similarities [11, 12, 42]. In video domain, the content data can be represented by different features, described with the following hierarchical levels: High-level features, representing *semantics* illustrated by the concepts and events happening within a video. An example can be a plot of the film *The Good, the Bad and the Ugly*, which showing three gunslingers who are competing to find a buried cache of gold during the American Civil War. Mid-level features, representing *syntactic features* with the existing objects within a video and interactions of these objects with each other. An example is people, horses and guns in the same film. At the lowest level, Low-level features typically representing *stylistic* aspects of videos defined by the aesthetic characteristics of the videos. This includes the design aspects which could picture the specific style of a video production. As an example, in the same movie predominant colors are yellow and brown.

Traditionally, content-based video recommendation has been focused on exploiting high-level and mid-level features [4, 22]. While they are effective in representing videos, however they are expensive to acquire as they need human-annotation typically by a large network of users. Indeed, there are cases where such features are missing and causing an *extreme* cases of cold start problem.

In such cases, even the most complicated recommendation algorithms could be unable to generate recommendation of such new items. In video domain, this is a case where a video is uploaded to a video-sharing platform and none of the users has yet added any type of data (e.g., tags).

In this paper, we address this problem and propose a novel feature set called *visual tags*. Such features are automatically extracted and added to the video items. We build predictive models that can learn the correlation among visual features and the tags added to other videos. Such models are used to predict the tags for a new video item with no tags. Such visual tags are then being exploited in order to generate personalized recommendation for users. We have performed different experiments in order to evaluate the recommendation based on visual tags. We have considered two evaluation scenarios, i.e., moderate cold start and extreme cold start scenarios. The results have shown the effectiveness of the proposed features in both scenarios in comparison to recommendation based on human-annotated tags.

It is worth noting that, we focused on recommendation based on tags as prior studies have shown the superior performance of tags

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

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

in comparison to other types of content-features (e.g., genre) [9]. Furthermore, using visual tags enables the system to include explanation when presenting recommended videos to users. Explanation may enhance transparency of the system and result in higher user satisfaction [37]. This is not very feasible with the pure (low-level) visual features.

2 BACKGROUND

This work is mainly related to two research fields, i.e., tag-based RSs and visually-aware RSs [5, 6, 22]. Several prior works have incorporated human-annotated tags into recommendation process [1, 16, 17, 27, 29, 40, 41]. One of the prior works [21] integrated tag-based similarity within an extended Collaborative Filtering (CF) in order to improve the recommendation. In [14], the authors proposed a modified version of the SVD++ matrix-factorization model [20] by replacing the usage of implicit feedback with tagging information. This results in a substantial improvement of the RS performance. In [24] another matrix-factorization model was proposed which expands the item model with latent factors vectors associated to the features of the items. In [15], SVD++ is again extended with an approach similar to those described in [14, 24], in order to deal with a cross-domain recommendation scenario.

The usage of the low-level visual features has drawn minor attention in RS (e.g., in [8, 25, 31]). This is while this has been extensively investigated in the other fields such as computer vision [28, 35]. [2, 19] provide comprehensive surveys on the state-of-the-art techniques related to video content analysis and classification, and discuss a large number of low-level features (e.g. visual, textual, or auditory). Authors in [30] propose a framework for movie genre classification based only on visual features. Moreover, [36] proposes a deep learning approach to automatically detect the director of a movie based on low-level visual features. This work differs from the prior works as it proposes *visual* tags instead of pure visual features. The advantage of using visual tags is the explainability of recommendation based on visual tags in comparison to visual features.

3 PROPOSED METHOD

Through querying YouTube, we obtained a huge dataset of 13923 movie trailers [7, 26, 31] based on the titles available in the Movielens dataset [18]. Prior work showed a high similarity of visual features extracted from movie trailers and their respective full-length movies [8].

Our research methodology encompasses the following steps: *Movie Segmentation*: every movie is segmented into *shots*, i.e., sequences of consecutive frames captured without interruption of the camera; *Key-Frame Detection*: within every shot the middle frame is selected as representative of the shot (key-frame); *Feature Extraction*: every key-frame is analyzed and the visual features are extracted; *Feature Aggregation*: the feature vectors are aggregated over the entire movie to form a feature vector descriptive of the whole movie; *Prediction*: aggregated visual feature vectors are used to train the prediction algorithms.

Movie Segmentation & Key-frame Detection. 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*. This is due to the fact that transition between two shots of the video is typically very abrupt. By comparing the color histogram of every movie frame, the histogram intersection is computed to compare the activities. Lets denote h_t and h_{t+1} as histograms of successive frames, then intersection is computed according to the following equation

$$\mathbf{s}(h_t, h_{t+1}) = \sum_b \min(h_t(b), h_{t+1}(b)) \quad (1)$$

where b is the index of the histogram bin. By comparing $\mathbf{s}$ with a predefined threshold (i.e., 0.75 in our experiment), we are able to segment the movies down to the constructing shots. Within every shot, the middle frame is selected as the key-frame.

Visual Feature Extraction. We have extracted a set of visual features capable of effectively capturing the *attractiveness* of each (key) frame of the movies. San Pedro and Siersdorfer [33] used a similar features set to predict popularity of Flicker images. In detail, 10 features are the following [26]:

- *Sharpness:* measures the clarity and level of details within the elements of a frame. This feature is related to the brightness contrast of edges in a frame.
- *Sharpness Variation:* is calculated via the standard deviation of all pixel sharpness values.
- *Contrast:* measures the relative difference in brightness or color of local features in a frame. Contrast is typically defined as the "assessment of the difference in appearance of 2 or more parts of a field seen simultaneously or successively". The root mean square contrast (RMS-contrast) is often used to compare frames [33].
- *RGB Contrast:* is almost identical to the basic contrast feature, explained before. However, it is extended to the three-dimensional RGB color space.
- *Saturation:* measures the colorfulness of the frame relative to the brightness. In the HSV color space the saturation estimation can be calculated via the RGB approximation of
$$frame\_saturation = \frac{1}{N}\sum_{x,y} S_{xy}, \text{ with} \quad (2)$$
$$S_{xy} = max(R_{xy}, G_{xy}, B_{xy}) - min(R_{xy}, G_{xy}, B_{xy})$$
where N is the amount of pixels in a frame and R_{xy}, G_{xy} and B_{xy} are the coordinates of the color of the pixel in sRGB space.
- *Saturation variation:* measures the variation in saturation via the sample standard deviation of all pixel saturations in a frame.
- *Brightness:* measures the average brightness of a frame; It uses a standard luminance algorithm
$$frame\_brightness = \frac{1}{N}\sum_{x,y} Y_{xy}, \text{ with} \quad (3)$$
$$Y_{xy} = (0.299 * R_{xy} + 0.587 * G_{xy} + 0.114 * B_{xy})$$
where Y_{xy} denotes the luminance value and N is the amount of pixels in a frame. R_{xy}, G_{xy} and B_{xy} are three RGB color space channels of pixel(x,y).
- *Colorfulness:* measures the individual color distance of the pixels in a frame. Therefore, the frame needs to be transferred into sRGB color space using $rg = R - G$ and $yb = 1/2\ (R + G) - B$.
- *Entropy:* of a frame is often used to determine how much information needs to be encoded by a compression algorithm. As an example, a frame with illustrating the moon craters has a very

high edge contrast, which leads to a high entropy. This means that the frame cannot be compressed very well which suggests that it can be used to measure the frame's texture.

- *Naturalness:* measures the difference (or similarity) between a frame and the human visual perception of the real world, with respect to colorfulness and dynamic range. Although subjective, it is an important visual quality metric when it comes to design [33]. We transfer the frame color space, if not already, to HSL. Then we use only pixels within the thresholds $20 \leq L \leq 80$ and $S \geq 0.1$. In the next step, pixels are grouped into one of the three sets 'Skin', 'Grass' or 'Sky', based on their H coordinate (hue). In order to calculate the naturalness of each set, the average saturation value of the group (μ_S) is used.

Feature Aggregation. To form the feature vector description of a movie, we have aggregated the visual features extracted from its key-frames. We have performed various aggregation functions, namely *median, standard deviation,* as well as, *1st & 3rd quartiles* of each visual feature, across all the key-frames of a movie. The last movie feature is the *number of key-frames*, within every movie. This process results in a vector of the length *41 aggregated features* per movie.

Recommendation algorithm We adopted a classical "K-Nearest Neighbor" content-based algorithm. Given a set of users $u \in U$ and a catalogue of items $i \in I$, a set of preference scores r_{ui} given by user u to item i has been collected. Moreover, 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*:

$$s_{ij} = \frac{\boldsymbol{f_i}^T \boldsymbol{f_j}}{\boldsymbol{f_i} \boldsymbol{f_j}} \tag{4}$$

For each item i the set of its nearest neighbors NN_i is built, $|NN_i| < K$. Then, for each user $u \in U$, the predicted preference score $\hat{r}_{ui}$ for an unseen item i is computed as follows

$$\hat{r}_{ui} = \frac{\sum_{j \in NN_i, r_{uj} > 0} r_{uj} s_{ij}}{\sum_{j \in NN_i, r_{uj} > 0} s_{ij}} \tag{5}$$

4 RESULTS

4.1 Experiment A: Analysis of Visual Features

After extraction of visual features, we discretized the features by decomposing them into three classes, i.e., (i) bottom class where the feature value is lower than the 1st quartile, (ii) middle class where the feature value is within the 1st and 3rd quartile, and (iii) top class where the feature value is higher than the 3rd quartile. Then we go over the features and for a given feature *f*, compute the percentage of movies belonging to each one of three classes of that feature. This results in a descriptive vector for each video allowing us to link tags to the visual features.

Furthermore, for the aim of reducing the data space, we use a powerful dimensionality reduction method called *T-distributed Stochastic Neighbor Embedding method (t-SNE)* [23]. The result is plotted in Figure 1. Please note that, every point in this figure represents a tag. As shown in the figure, the tags could be positioned close to or far from each other, depending on their visual similarity. Our observation shows that, although the similarity is computed based on low-level visual features that are not really semantic,

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

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-middle side of the figure are semantically related: *murder, mafia, police, heist,* and *crime.*

4.2 Experiment B: Cold Start Recommendation

We used the *visual* tags extracted from videos based on low-level visual features to build a content-based RS. We evaluate the system under the *new item* cold start situation. In the extreme cold start, when a new video has been added to system catalogue with no rating and no content features (e.g., tags) provided by users, only visual tags and visual features can be used as they don not need any human-annotation. In addition to that, we considered also the moderate cold start where the new item has received no rating but a few tags from users, that can be used to generate content-based recommendation. We compared the performance of recommendation based on (automatic) visual tags and (automatic) visual features, against (human-annotated) tags when the number of tags is increased from 1 to 10, in terms of prediction accuracy (i.e., MAE and RMSE) and Coverage [34]. We 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) [9].

Figure 2 (left) shows the results in terms of MAE. As it can be seen, in the early phase of the cold start, by far the best result has been achieved by recommendation based on visual features and visual tags with lowest MAE values. However, when sufficient number of tags are collected (7 tags), tag-based recommendation overtakes recommendation based on these features.

Similar results have been observed for the RMSE metric. Figure 2 (middle) presents the results for RMSE. As it can be seen, again, in the early stage of the cold start, the recommendation based on visual features and visual tags express substantially better performance by achieving the lowest RMSE values. This is the case up to when 8 tags are collected by the system for the video items.

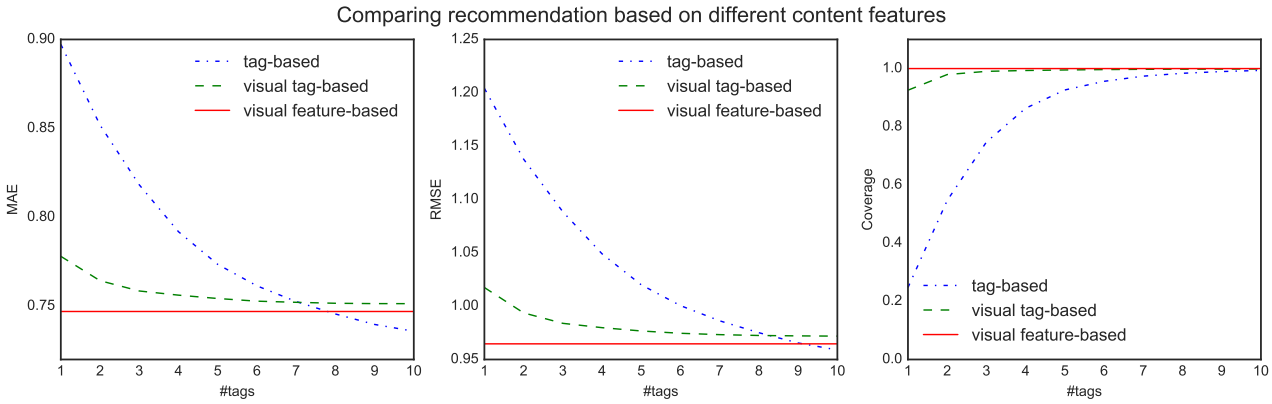

Figure 2: Comparing recommendation based on different features, in terms of: (left) MAE, (middle) RMSE, and (right) Coverage.

However, as soon as 9 tags (and more) are collected, the tag-based recommendation outperforms the recommendation based on other features by achieving a lower RMSE value.

This is an interesting outcome and it shows that in the extreme cases of cold start situation, when no or very few tags are available, (classical) tag-based recommendation may fail to predict the true preferences of the users and properly generate relevant recommendation for users. However, in the moderate cold start, when larger number of tags are collected from users, the tag-based recommendation overtakes recommendation based on visual tags and features.

We have also compared the performance of recommendation based on these content features, in terms of Coverage, i.e., percentage of ratings that recommender can predict over all missing and available ratings. Figure 2 (right) shows the obtained results. In terms of coverage, the best results have been achieved by visual features and visual tags. Due to the nature of the visual features, represented as a full matrix with no sparsity (see Figure **??**), recommendation based on visual features is capable of predicting the entire ratings and achieving 100% coverage, even in the extreme cold start situation. Visual tags can achieve similar performance with automatically generating 5 and more visual tags. This is possible for human-annotated tags after collecting 9 and more tags.

Overall, these promising results illustrate the significant power of visual tags and visual feature, with no need for costly human-annotation, in dealing with the extreme and moderate cases of new item cold start problem.

5 CONCLUSIONS & FUTURE WORK

This paper addresses the cold start problem by proposing recommendation based on *visual* tags. These are novel form of content features that can be automatically made for new items, where neither any rating nor any tag is available for that item. We have conducted a preliminary experiments to better understand the potential correlations among visual features and user tags. This help us to better model and implement the mechanism to extract visual tags. In addition to that, we have evaluated the quality of recommendations based on visual tags and compared it to the actual tags collected from users in the cold start situation. The observed results are promising and show the power of visual features in dealing with even the extreme cases of cold start problem.

In future, we will elicit user-generated video content from other video sharing social networks (e.g., Instagram). We also plan to integrate and put a new layer over our model using the recent innovations [38, 39] in order to obtain the implicit preferences of users through their facial appearance. We will use noted innovations in order to study the potential correlation between users' facial expressions versus the movie features. We will also explore the potential of novel interaction design when developing a visually-aware movie recommender system [3, 10].

REFERENCES

[1] Toine Bogers. 2018. Tag-based recommendation. In *Social Information Access*. Springer, 441–479.
[2] 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
[3] Paolo Cremonesi, Mehdi Elahi, and Franca Garzotto. 2017. User interface patterns in recommendation-empowered content intensive multimedia applications. *Multimedia Tools and Applications* 76, 4 (01 Feb 2017), 5275–5309. https://doi.org/10.1007/s11042-016-3946-5
[4] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What videos are similar with you? learning a common attributed representation for video recommendation. In *Proceedings of the 22nd ACM international conference on Multimedia*. 597–606.
[5] Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. Semantics-Aware Content-Based Recommender Systems. In *Recommender Systems Handbook*. Springer, 119–159. https://doi.org/10.1007/978-1-4899-7637-6_4
[6] Marco Degemmis, Pasquale Lops, and Giovanni Semeraro. 2007. A Content-collaborative Recommender That Exploits WordNet-based User Profiles for Neighborhood Formation. *User Modeling and User-Adapted Interaction* 17, 3 (July 2007), 217–255. https://doi.org/10.1007/s11257-006-9023-4
[7] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, and Pietro Piazzolla. 2016. Recommending Movies Based on Mise-En-Scene Design *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1540–1547. https://doi.org/10.1145/2851581.2892551
[8] 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
[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* 7, 4 (2018), 207–219. https://doi.org/10.1007/s13735-018-0155-1
[10] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, Paolo Cremonesi, and Franca Garzotto. 2015. Toward Effective Movie Recommendations Based on Mise-en-Scène Film Styles. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*. ACM, 162–165.

[11] Farshad Bakhshandegan Moghaddam1 Mehdi Elahi et al. 2019. Cold Start Solutions For Recommendation Systems. IET.

[12] 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.

[13] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review* 20 (2016), 29–50.

[14] 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

[15] 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.

[16] Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu. 2007. The benefit of using tag-based profiles. In *2007 Latin American Web Conference (LA-WEB 2007)*. IEEE, 32–41.

[17] Mouzhi Ge, Mehdi Elahi, Ignacio Fernaández-Tobías, Francesco Ricci, and David Massimo. 2015. Using Tags and Latent Factors in a Food Recommender System *(DH '15)*. ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/2750511.2750528

[18] 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

[19] 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

[20] 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

[21] 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

[22] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul Kantor (Eds.). Springer, 73–105.

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

[24] 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

[25] 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

[26] Farshad B Moghaddam, Mehdi Elahi, Reza Hosseini, Christoph Trattner, and Marko Tkalčič. 2019. Predicting Movie Popularity and Ratings with Visual Features. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. IEEE, 1–6. https://doi.org/10.1109/SMAP.2019.8864912

[27] Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, and Shunsuke Uemura. 2007. Tag-based contextual collaborative filtering. In *Proceedings of the 18th IEICE Data Engineering Workshop*. 377–386.

[28] 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

[29] Zhou Qingbiao, Fang Jie, and Gaundong Xu. 2011. Incorporating sentiment analysis for improved tag-based recommendation. In *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 1222–1227.

[30] 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

[31] 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

[32] Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan. 2015. Active Learning in Recommender Systems. In *Recommender Systems Handbook*. Springer, 809–846.

[33] 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

[34] 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.

[35] 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

[36] 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

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

[38] 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

[39] 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

[40] Christian Wartena, Rogier Brussee, and Martin Wibbels. 2009. Using tag co-occurrence for recommendation. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, 273–278.

[41] Deqing Yang, Yanghua Xiao, Yangqiu Song, Junjun Zhang, Kezun Zhang, and Wei Wang. 2014. Tag propagation based recommendation across diverse social media. In *Proceedings of the 23rd International Conference on World Wide Web*. 407–408.

[42] Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang, and Tao Zhou. 2010. Solving the cold-start problem in recommender systems with social tags. *EPL (Europhysics Letters)* 92, 2 (2010), 28002.