Politecnico di Milano
Department DEIB
Doctoral Programme In INFORMATION TECHNOLOGY

Video Recommendation by Exploiting the Multimedia Content

Doctoral Dissertation of:
Yashar Deldjoo

Supervisor:
Prof. Paolo Cremonesi

Tutor:
Prof. Andrea Bonarini

The Chair of the Doctoral Program:
Prof. Andrea Bonarini

2018 – Cycle XXX

This work has been funded by TIM S.p.A., Services Innovation Department, Joint Open Lab Milano, Italy.

I dedicate this thesis to **my dear mother** and **my dear father** to whom I owe everything I have and to the love of my life **Atena** ...

Contents

1 Introduction 1

1.1 Motivation . . . 1

1.2 Research Goals . . . 3

1.3 Contributions . . . 7

1.4 List of Publications . . . 10

1.5 Outline of the Thesis . . . 12

2 Recommendation using Multimedia Content Analysis: State of the Art and Preliminaries 13

2.1 Foundations of multimedia processing and recommender systems . . . 15

2.1.1 Multimedia processing . . . 16

2.1.2 General framework . . . 32

2.2 State-of-the-Art Approaches . . . 40

2.2.1 Improve recommendation quality: . . . 41

2.2.2 Popularity bias and long tail . . . 43

2.2.3 Multimodal analysis . . . 46

2.2.4 Context-aware multimedia recommendation . . . 50

2.2.5 Leveraging social network information . . . 61

2.2.6 Sequence-aware multimedia recommendation: . . . 62

3 Visual Content-Based Video Recommendation 67

3.1 Introduction . . . 67

3.2 Study 1: "The impact of low-level visual features" . . . 70

3.2.1 Background and Related Work 70
3.2.2 Method Description 71
3.2.3 Evaluation Methodology 75
3.2.4 Results 76
3.2.5 Feature Analysis 80
3.2.6 Discussion 83
3.2.7 Conclusion of Study 1 84
3.3 Study 2: "User study" 85
3.3.1 Introduction and Context 85
3.3.2 Experiment Setup 87
3.3.3 Study A: Offline Experiment 88
3.3.4 Study B: Online Evaluation 89
3.3.5 Experiments and Results 92
3.3.6 Discussion 94
3.3.7 Conclusions of study 2 95
3.4 Study 3: "Hybridization with semantic-rich systems" 95
3.4.1 Introduction and related work 96
3.4.2 Hybridization using Canonical Correlation Analysis . 97
3.4.3 Hybridization using Factorization Machines 100
3.4.4 Experimental Results for study A and B 101
3.4.5 Conclusion of study 3 105
3.5 Conclusion of Chapter 3 105

4 Multimodal Content-Based Video Recommendation 107

4.1 Introduction 107
4.2 Related Work 110
4.2.1 Multimedia Recommender Systems 110
4.2.2 Multimedia Content Description 112
4.2.3 Contributions 114
4.3 Proposed Recommendation Scheme 115
4.4 Rich Item Descriptions Defining the Video Genome 116
4.4.1 Audio Features 116
4.4.2 Visual Features 125
4.5 Evaluation Setup 129
4.5.1 Content Descriptors 129
4.5.2 Recommendation Algorithm 131
4.5.3 Hybridization 131
4.6 Experimental Study A: Offline Experiment with Simulated Users 133
4.6.1 Data 133

4.6.2 Objective Evaluation Metrics 133
4.6.3 Recommendation *w.r.t.* Accuracy in Cold-Start Scenario 136
4.6.4 Recommendation *w.r.t* Accuracy in Warm-Start Scenario 139
4.6.5 Recommendation *w.r.t* Beyond Accuracy Metrics . . 140
4.7 Experimental Study B: Online Evaluation 144
4.7.1 Perceived Quality Metrics 144
4.7.2 Evaluation Protocol 145
4.7.3 Results 147
4.8 Experimental Study C: Addressing the Open Research Questions 150
4.8.1 RQ1: *video genome* vs. metadata for video recommendation 150
4.8.2 RQ2: which audio-visual information is better 151
4.8.3 RQ3: is fusing information more effective 151
4.9 Conclusions of Chapter 4 151

5 Evaluation and Dataset 155
5.1 Movie Datasets 155
5.2 Previous Work 156
5.3 Dataset Description 159
5.3.1 Provided content descriptors 159
5.3.2 Dataset basic statistics 161
5.3.3 Data format 161
5.4 Ground Truth 164
5.5 Baseline and Reference Results 166
5.6 Evaluating video and music recommender systems 166
5.6.1 Accuracy metrics 167
5.6.2 Beyond-Accuracy metrics 170
5.6.3 Sequence-Aware multimedia evaluation metrics . . . 171
5.7 Conclusions of Chapter 5 172

6 Conclusion 173
6.1 Summary of Thesis 173
6.2 Main Contributions 173
6.2.1 Characterizing recommender systems exploiting multimedia content 174
6.2.2 Video recommendation by exploiting the visual content 174
6.2.3 Movie trailer v.s. full movies 175

6.2.4 Video recommendation by exploiting the latest advances in audio and visual content analysis, both in offline experiments and via user-studies in a real deployed system 176
6.2.5 Present a novel large-scale dataset to facilitate research on movie recommendation and retrieval 177
6.3 Outlook and Future Works 177

Bibliography **179**

Bibliography **179**

Acknowledgements

First and foremost, I would like to express my gratitude and appreciation to my advisor and mentor Paolo Cremonesi. In particular, I highly appreciate his scientific excellence, his love of personality, his efforts to establish a friendly working environment and his support whenever I needed it. I would like to thank Paolo especially for building the trust on me and giving me the freedom to perform interesting research, according to my own interests and background at the time I was a master student. Paolo's advices have helped me a lot to enter a productive line of research, considering the time constraints during the Phd programme in a competitive environment of Politecnico di Milano, one of the most excellent technical universities in Italy. I thank him for helping me writing the papers specially during the beginning of my PhD programme and to let me enter the community of Recommender Systems in a relatively easier manner. I thank him for his help on improving my papers even if it required staying up late before a submission deadline. Finally, I would like to thank him for his positive attitude for me to follow an abroad visiting period to Johannes Kepler University during the course of Phd, even if this was not mandatory, which came to be very useful for advancing the research field I have been pursuing.

I would also like to thank my co-advisor Markus Schedl for hosting me at Department of Computational Perception, Johannes Kepler University (in Linz, Austria), for his endless motivation, energy, immense experience and for many interesting discussions and collaborations we have had which are still on going. Having an invaluable experience and knowledge in both fields of Multimedia and Recommender System, I found the opportunity of working with Markus very useful. Despite the self-determined and productive line of research we are collaborating on, Markus's friendly behavior helps a lot to establish a stimulating and productive research collaboration environment. I would like to thank all members at JKU for being so friendly and for the outstanding excellence of the department which performs cutting-edge research in the field of machine learning and music information retrieval.

I would like also to express my gratitude to Prof. Gabriella Pasi at university of Milanio-Bicoca which provided me a financial support in the final months of my PhD thesis effectively allowing my thesis to be completed in an extremely peaceful and relaxing manner and initiated a successful research collaboration in the follow-up.

Furthermore, I wish to express my gratitude to all the great people with whom I had the pleasure to collaborate during the past few years, most notably Mehdi Elahi my long-lasting classmate and closest friend who had the main role in introducing me to the PhD programme at Politecnico di Milano and recommending me to my adviser. I had the pleasure to collaborate with my old friend Mehdi during the Phd programme for about 2 years. He has encouraged me a lot, supported me through the difficult times and celebrated with me through the good times. Having an amazing friend like Mehdi is the best luck one can have in his lifetime. I would also like to thank my other great friends Massimo Quadrana for the collaboration we have had and Roberto Pagano the other colleague of mine for their amazing friendliness.

I would also like to acknowledge TIM S.p.A., Services Innovation Department, Joint Open Lab Milano, Italy, mainly Massimo Valla and Cristina Fra for the financial support of my Phd research, providing and excellent environment for pursuing research and for some of the interesting discussions that we have had together. Without their support, this Phd would have never come to an end. Massimo and Cristina I am greatly thankful to both of you and Telecom Italia.

Last but not the least, I would like to thank my entire family in Iran and my beautiful wife Atena who supported me, inspired me, and loved me. I would like to thank sincerely my **very kind and supportive father** for being the main source of hope, support and inspiration on me to cultivate the thirst for learning in my mind from early years of childhood, for **the most kind-hearted mother** whose supports and love for me are endless in my life and I owe her everything for who I am, **my great brother** who has always been supportive of our family and finally **my beautiful wife Atena**, the love of my life for her being a great partner and supporter of me in times of comfort and difficulty. I would also like to thank my **kind mother-in-law** for her patience and support of my family at the time we were far from her. My marriage with Atena and all our first memories coincide with my PhD period and I think I will never forget this period of my life. I will always love you **Atena** and will be always with you. You are the greatest gift of God to me! (**Dooset daram** = I love you in Farsi)

Abstract

VIDEO recordings are complex media types. For example, when we watch a movie, we can effortlessly register a lot of details conveyed to us (by the author) through different multimedia channels, in particular, the audio and visual channels. To date, the majority of *content-based movie recommender systems* (CBMRS) base their recommendations on metadata (*e.g.,* editorial metadata such as genre or wisdom of the crowd such as user-generated tags) since they are human-generated and are assumed to cover the 'content semantics' of movies by a great degree. Multimedia features, on the other hand, provide the means to identify videos that 'look similar' or 'sound similar'. These discerning characteristics of heterogeneous feature sets meet users' differing information needs.

In the context of this PhD thesis, methods for automatically extracting video-related information from the multimedia content (*i.e.,* audio and visual channels) have been elaborated, implemented, and analyzed. Novel techniques have been developed as well as existing ones refined in order to extract useful information from the video content and incorporate them in recommendation systems. Different video recommendation tasks are solved using the extracted multimedia information under recommendation models based on content-based filtering (CBF) models and the ones based on combination of CBF and collaborative filtering (CF).

As a branch of recommender systems, this thesis investigates a particular area in the design space of recommender system algorithm in which the generic recommender algorithm needs to be optimized in order to use a wealth of information encoded in the actual image and audio signals.

The results and main findings of these assessments are reported via several offline studies or user-studies involving real users testing a prototype of developed movie recommender systems powered by multimedia content. The results are promising and show different scenarios in which multimedia content can be leveraged for successful video recommendation outperforming the alternatives, most notably in new-item settings.

CHAPTER 1

Introduction

1.1 Motivation

In recent years, we have witnessed a rapid growth in availability of many forms of video content (*e.g.,* user-generated videos, movies, music video clips) created, shared, and consumed through various web channels such as YouTube [1], Netflix [2], DailyMotion [3], Vimeo [4], *etc.,* as well as the social media platforms. According to statistics, as of 2016, more than 500 hours of videos were uploaded to the channel every minute; this is equivalent to say if one would like to watch all the videos uploaded to YouTube only *in one hour*, it would take a genuine time of 3 years of uninterrupted watching to see them all [4]. Another statistics by Cisco, the largest network company in the world, reveals that more than 75% of world's mobile data traffic will be video by the end of 2020 and (more than) 80% when video and audio data are considered together [1]. When faced with this amount of information, consumers often feel overwhelmed. How much information can people reasonably process? Even retrieving a specific version of an item amidst

[1] https://www.youtube.com

[2] https://www.netflix.com/it-en/

[3] http://www.dailymotion.com

[4] https://www.vimeo.com

pages and pages of Internet is sometimes equal to looking for the proverbial "needle in the haystack" [40]. As a countermeasure, recommender systems (RS) that help users with decision-making on which products to purchase/-consume by automatically determining the content that a user may like have emerged and developed during the last decade [16, 23, 268].

Video recordings are very complex signals composed of different multimedia channels, in particular, the audio and visual modalities. When we watch a video, we can effortlessly register a lot of details communicated to us through these modalities. As a result, the video content can be described in versatile manners since its perception is not limited to one sense. For example, for a movie, these multiple facets could be reflected in its genre, its reviews, and its visual and audio content. To date, metadata features are commonly used in the design of video and other multimedia recommender systems (*e.g.,* music RS) as they are assumed to cover the 'content semantics' of a video, mainly because metadata is the result of human knowledge, reflecting either expert knowledge in case of editorial metadata like genre, or the wisdom of the crowd in case of user-generated tags/keywords. Multimedia information, on the other hand, provide the means to identify videos that 'look similar' or 'sound similar'. These discerning characteristic of multimedia meet users differing information needs.

Video recommendation systems are traditionally powered by either collaborative filtering (CF) or content-based filtering (CBF) algorithms [190, 214, 268]. CF exploits the *interactions* between users and items to make recommendations. In other words, CF methods leverage the fact that the specified ratings are often highly correlated across users and items, therefore the unspecified ratings (to be predicted by the RS) can be imputed by considering the *inter-item* correlations (*item-based* CF model) or *inter-user* correlations (*user-based* CF model). Some models use both types of correlations [16, 23]. CBF approaches on the other hand, use content descriptors of items expressed in terms of attributes, typically feature vectors, to make recommendations. In situations when a new item is added to the catalog and not much rating is available about the item (new item problem), CBF are preferred over CF since CBF models do not require preference of the other users and as long as some piece of information about the target-user's *own preference* are available, recommendations can be computed.

The extent to which CBF and their respective algorithms are used strongly varies between domains, *i.e.,* the type of item being recommended. While in the multimedia community, extracting descriptive item features from the multimedia content (*i.e.,* text, audio, image, and video content) is a well-researched task, the recommender systems community has considered for a long time metadata, such as title, genre, tags, actors, or plot of a movie, as the single source for CBF recommendation models, thereby disregarding the wealth of information encoded in the actual content signals. In addition, the metadata generations process is a time-consuming and error-prone task (in particular being user-generated) and labor-intensive/expensive to collect (especially the expert-generated ones). User-generated metadata often exhibit user or community biases and might therefore not fully, or only in a distorted way, reflect the characteristics of a video [61, 197]. Furthermore, metadata are often rare or absent for new videos, making it difficult or even impossible to provide good quality recommendations — a problem known as *cold-start scenario* [269][5]. Even if metadata are available in abundance, given their unstructured or semi-structured nature, they often require complex natural language processing (NLP) techniques for pre-processing, *e.g.,* syntactic and semantic analysis, stemming, or topic modeling [19].

Addressing theses research gaps, in this thesis I explore solving different video recommendation tasks using multimedia content. As a branch of recommender systems, the thesis investigates a particular area in the design space of recommender system algorithm in which the generic recommender algorithm needs to be optimized in order to use a wealth of information encoded in the actual image and audio signals. I believe that recommender system research can strongly benefit from knowledge in multimedia signal processing (MSP) and machine learning (ML) established over the last decades for solving various multimedia processing and retrieval tasks. In this thesis, I aim to bridge the gap in perspectives and advances between the MSP/ML and the RS communities. When I started this thesis using visual features for video recommendation showed some initial success for video recommendation both in addressing the cold-start scenarios and improving the performance in warm-start scenarios; however the topic was heavily under-researched. Feeling this gap seemed promising for advancing the field of video recommendation and in general multimedia recommendation but also for learning about multimedia processing and subsequently transferring the expertise and domain knowledge required to apply to video signals.

1.2 Research Goals

In this thesis, we explore to solve different video recommendation problems by exploiting the visual and audio content of videos. A broad range of

[5] Cold-start refers to the situation where rating and/or metadata are rare or absent. Here we use the terminology to refer to the lack of metadata as CBRS do not require the ratings of other users to generate recommendations.

research questions are presented throughout chapters of this thesis. The main research goals addressed are as following:

RG1: Review the state of the art on recommender systems exploiting multimedia content analysis: We aim to present a systematic literature review of research works exploiting multimedia content for a particular media content recommendation, such as music (example of audio media), movie (example of video media) or clothes in fashion industry (example of image media type) [226] among others and also in domains where the target product is not necessarily a media type but some form of information presented to the user, for example in tourism domain recommending place of interest (POI) to users by analyzing visual content of user-generated photos [71]. The literature-review aims to provide an extensive and rigorous discussion of the state of the art and latest advances in the field in order to bridge the advances in the field of recommender systems and multimedia processing.

RG2: Investigate the impact of personalized video recommendations using the visual content

From the survey in RG1, we realized that multimedia content has been widely adopted in a few neighboring fields to video RS such as music information retrieval (MIR) to solve variety of music-related tasks including music information retrieval, music recommender system and music similarity estimation. This is while the number of research works dealing with the video recommendation problem using the multimedia content is by a large margin lower where this can have several reasons, complexity of video signal compared with music with regards to different modalities involved in it, the duration of signal, the level of user emotional and cognitive engagement when watching movies compared with music and the absence of repeated recommendation nature in movies compared with music to name a few. When it comes to images, based on our literature review in RG1, we believe the number of research articles dealing with images (*e.g.,* consider [66, 71, 149, 226, 348]) is larger than videos but still less than those in music domain . We therefore focused our attention mainly on the video domain and techniques that can be applied to advance the field of video recommendation by exploiting multimedia content.

Given that the name of the video is intrinsically linked with the visual modality (similar to audio modality for music), in RG2 we base our goal on investigating if the visual characteristics of video contains useful information that could be leveraged in order to improve the utility of recommendation? Which visual content is the key and how recommendations quality is changed when building CB or hybrid CB+CF video recommendation systems that leverage visual content for predicting user's preference?

Having studied the literature of multimedia processing (and computer vision) in the preliminary phases of the research, we identified a number of visual features based on color, motion and lighting that showed initial success in beating simple CB recommendation based on K-nearest neighbor (KNN) using genre as metadata. Once we verified the effectiveness of the recommendations using the visual content, we pursued to advance the field by improving visual features, techniques for representation the features, techniques for fusing this source of information with other available data (*e.g.,* metadata or rating) techniques as well the core recommendation method.

RG3: Investigate the impact of perceived personalized video recommendations using the visual content, metadata and combination of both

From the analysis in RG1 and RG2, we realized that majority of the existing research works focous on content-based recommendation by exploiting multimedia content for predicting rating scores of users given to items (rating prediction task) or recommending the best K items matching the short-term and long-term preference of users (top-K recommendation problem). These works employ variety of accuracy metrics such as: error metrics (RMSE, MAE), decision-support metrics (precision, recall, F1), or rank-aware metrics (MAP, MRR, NDCG). A few works extend the accuracy metrics and consider beyond-accuracy metrics (novelty, diversity, catalog coverage) as the secondary goal of recommendation system. All these studies are carried out using historical dataset via the so-called offline experiments. However, how users perceive the quality of recommendation in a real system can be different. We noticed an overall lack of user-studies in movie recommendation exploiting multimedia content while in music domain, a considerable amount of research works of this type exists.

Therefore, once we assessed the effectiveness of the recommendations in offline experiment, we pursued deploying the algorithm on a real-system in order to evaluate if the introduction of low-level visual features alone or combined with high-level semantic features, induce measurable effects on the *perceived* utility of recommendations.

RG4: Trailers *v.s.* Full which content is key? Movies and the corresponding trailers are not made necessarily by the same director. In the former case, the director applies various filmmaking conventions to influence the believability of a film in the eyes of its viewers. These conventions are sometimes referred to as "mise-en-scene". Elements of the mise-en-scene include but not limited to composition, costumes, actors, objects, movements, lighting, and sound or everything that appear in front of the camera and its arrangements. While full-length movies are made with a natural pace to increase the believability of the scenes, trailers (aka previews) serve as an advertisement for a feature film that is planned to be exhibited on TV or at the cinema. Movies trailers are nowadays popular on the Internet and mobile device and particular contextual situations like an air-flight in order to give the user an option for "previewing" the movies before actually watching them. As such, trailers can be made with a lot of abrupt cutting and merging of the scenes since their production goal is mainly to "excite the users" in order to watch the full movies, even if this comes at the expense of "eradicating" the mise-en-scene. The correspondence of the trailers with their full-movies can be dependent on the genre, country of their production and other factors.

RG5: Investigate the impact of latest state of the art audio and visual content on the quality of video recommendation both in terms of offline quality of recommendations and user-studies and releasing a new, large-scale video dataset

Based on the promising achievements obtained in RG2 and RG3 by using the visual characteristics of videos, in RG5 we aim to deploy a complete movie recommender system which employs the latest state of the art descriptors in the field of multimedia. RG5 improves and extends RG2 by large margin when considering the following aspects:

- In RG5, we consider both visual *and aural* characteristics to filter the movies while in RG2 we considered the singe visual modality. This is an important extension since the communication of emotions/messages from the director to the user is by nature multi-modal (see section 2.2.3 for further information) and a complete RS should consider both of these sources of information in meeting users' diverse information needs. This problem has been heavily unsearchable in the community, the only ones to date we can name of are [229, 339] using quite old deprecated features and posing the problem of recommendation as visual retrieval problem.

- We employed the latest state of the art audio and visual features in the proposed system in RG5, specifically the one based on i-vectors and convolutional deep neural networks, where the former has had a huge success for tasks related to speaker verification, music information retrieval, music similarity estimation and acoustic scene classification (all involved in movies) and the latter with a recognized performance for visual object recognition. The system also employed other types of established features such aesthetic visual features (used for measuring beauty in images) and block-level audio features for completeness.

- We proposed an effective hybridization (fusion) technique leverage the informativeness of different feature sets when combined together.

- The final model in a real movie recommender system similar to Netflix whose performances was validated in RG2 and published as a short paper at ACM conference on Recommender Systems [112]. All the experiments are therefore validated comparatively in an offline setting using historical dataset and via a user-study. In the offline setting, we validate the performance of the proposed CB models on a large-scale movie trailer dataset while with the user-study assessed the results obtained in the offline study by measuring the perceived utility of recommendation.

- Finally, to address the need for large datasets for movie recommendation, we aim to present a novel, open, large-scale dataset for devising and evaluating movie recommender systems.

1.3 Contributions

Over the past three and a half years, my principal intention has been bringing together the fields of recommender system and multimedia information retrieval/signal processing. The focus of this thesis is on video products (specifically movies) but the ideas learned can be easily mitigated to many neighboring domains (*e.g.,* images, music, Web) as well domains where the target product is not necessarily a media content (*e.g.,* social networks and tourism).

The scientific contributions from this endeavor are manifold:

1. *A comprehensive, in-depth review of the state of the art in recommender systems exploiting multimedia content analysis (discussed in Chapter 2*

   We aim to bridge the gap in perspectives and content analysis advances between the *multimedia* and the *recommender systems* communities.

   In section 2.1, we present **foundations of multimedia processing and recommender systems** and explain the basic concepts of multimedia signal processing and recommender systems, disambiguate the term multimedia as an item for recommendation and feature extraction modality, present different models of feature extraction techniques, and different types of content-based recommendation algorithms (and hybrid CB+CF ones) from basic K-nearest neighbor (KNN) method (leveraging solely content) to more complex ones such as *regression-based latent-factor model* (RLFM) or *factorization machines* (FM) (leveraging both content and user ratings).

In section 2.1.2, we present **the general framework for building a content-based RS exploiting the multimedia content** and describe the nuts and bolts of different processing steps involved in each stage for each modality.

In section 2.2, we present **an in-depth review of the state of the art in recommender systems exploiting multimedia content** and categorize the reviewed state-of-the-art research works with respect to the objectives and challenges they address.

2. *Video recommendation by exploiting the visual content (discussed in Chapter 3)*

   Based on my publications [87, 88, 90, 91, 94, 112], in this chapter we focus **on solving several video recommendation problems by exploiting low-level visual features** extracted from the visual channel of movies. I proposed a new type of CB technique that filters videos according to the visual characteristics of movies including color variation, motion and lighting key where these features correspond with the stylistic elements used in applied media aesthetics [344] used to convey communication effects and to simulate different feelings in the viewers. Several subproblems are addressed in this context including analyzing the quality of different low-level visual features in offline experiments in both cold-start and warm-start scenarios, investigating the perceived utility of recommendation using low-level visual features alone or hybridized with high-level semantic features, analyze the offline quality of recommendation under hybridization techniques based on canonical correlation analysis (CCA) and factorization machines (FM).

3. *Video recommendation by exploiting the latest advances in audio and visual content analysis, both in offline experiments and via user-studies in a real deployed system*

   Based on my recently submitted publication to the journal of User Modeling and User-Adapted Interaction (UMUAI), **I deployed a CB**

video RS that exploits latest state-of-the-art visual and audio features along with metadata, in order to build *rich item descriptions*. We refer to this rich content information as the *Video Genome*, since it can be considered as the footprint of both content and style of a video (similar to a biological DNA composed of long sequences of letters) [52]. I further proposed an improved version of Borda count method to hybridize the ranking output of different recommenders and show that under this hybridization, the quality of recommendation is significantly improved compared when the standard form of Borda count method or when the recommenders are used individually.

For evaluation, we conducted two large-scale empirical studies: (i) a system-centric study to measure the offline quality of recommendations in terms of accuracy-related metrics such as MRR, MAP, Recall and beyond-accuracy novelty, diversity, and coverage performance measures, and (ii) a user-centric online experiment, measuring different subjective metrics, including relevance, satisfaction, and diversity. experiments have shown, multimedia features can provide a good alternative to metadata with regards to both accuracy measures and beyond accuracy measures. The most significant improvement for the accuracy metrics was achieved when using novel, state of the art approaches for audio and visual features: i-vectors and deep neural network features. Multi-modal fusion under the proposed Borda count method shows the best results in our studies.

4. *Present a large-scale dataset for video recommendation using multimedia content and discuss the main evaluation metrics used to assess the quality of video recommendation algorithms*

   Based on my publication [86] (accepted in April 2018), I designed and released a large-scale publicly available multifaceted movie trailer dataset (MMTF-14K) to facilitate research on recommendation, classification and retrieval. MMTF-14K contains audio and visual descriptors in addition to ratings and metadata for 13,623 Hollywood-type movie trailers.

   With respect to existing datasets, this is first large-scale and stable dataset in the community of RS which supplied all types of precomputed content-based descriptors in conjunction with metadata in numerical feature format. Part of these data is used in a MediaEval 2018 task[6] (See [85] for more information).

[6] http://www.multimediaeval.org/mediaeval2018/

1.4 List of Publications

The following articles were published during the course of this research:

Under Review

Deldjoo, Y., Constantin, M.G., Eghbal-zadeh, H., Cereda, S., Schedl, M., Ionescu B., Cremonesi, P. (2017), *Video Genome: Audio-visual Encoding of Multimedia Content to Enhance Movie Recommendations*, Journal of User Modeling and User-Adapted Interaction (UMUAI), (submitted Dec 2017)

Journal Publications

Deldjoo, Y., Elahi, M., Cremonesi, P. and Quadrana, M. (2018), *Using Mise-en-Scène Visual Features based on MPEG-7 and Deep Learning for Movie Recommendation*, International Journal of Multimedia Information Retrieval (IJMIR).

Schedl, M., Zamani, H., Chen, C.W., **Deldjoo, Y.**, and Elahi, M. (2018), *Current Challenges and Visions in Music Recommender Systems Research*, International Journal of Multimedia Information Retrieval (IJMIR)

Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., and Quadrana, M. (2016), *Content-based video recommendation system based on stylistic visual features*, Journal of Data Semantics (JoDS), 5(2):99-113.

Conferences Publications

Deldjoo, Y., Constantin, M.G., Eghbal-zadeh, H., Cereda, S., Schedl, M., Ionescu B., Cremonesi, P. (2018), *MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval*, Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, Netherland, June 12-15, 2018

Elahi, M., **Deldjoo, Y.**, Moghaddam, F.B., Cella L., Cereda S., and Cremonesi P., *Exploring the semantic gap for movie recommendations*, Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, pp. 326-330, Como, Italy, August 27-31, 2017

Deldjoo, Y., Elahi, M., Cremonesi P., Moghaddam F.B., and Caielli A. L. E., *How to combine visual features with tags to improve movie recommen-*

dation accuracy?, in E-Commerce and Web Technologies: 17th International Conference, EC-Web 2016, pp. 34-45, Porto, Portugal, September 5-8, 2016

Deldjoo, Y., Elahi, M., Quadrana, M., Cremonesi P., and Garzotto, F., *Toward Effective Movie Recommendations Based on Mise-en-Scène Film Styles*, ACM Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter, CHItaly 2015, 162-165, Rome, Italy, September 28-30, 2015

Deldjoo, Y., Elahi, M., Quadrana, M., and Cremonesi P., and Piazzolla, P. *Toward building a content-based video recommendation system based on low-level features*, in E-Commerce and Web Technologies: 16th International Conference, EC-Web 2015, pp. 45-56, Valencia, Spain, September 2015

Workshops/Extended Abstracts

Deldjoo, Y., Cremonesi P., Schedl M., and Quadrana, M., *Enhancing Children's Experience with Recommendation Systems*, Workshop on Children and Recommender Systems (KidRec '17) as part of ACM Conference on Recommender Systems 2017, Como, Italy, August 27-31, 2017

Abel, F., **Deldjoo, Y.**, Elahi, M., and Kohlsdorf D., *Recsys challenge 2017: Offline and online evaluation*, Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, pp. 372-373, Como, Italy, 2017/08/27

Deldjoo, Y., Cremonesi P., Schedl M., and Quadrana, M., *The effect of different video summarization models on the quality of video recommendation based on low-level visual features*, ACM Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI 2017, pp. 20:1–20:6, Florence, Italy, June 19-21, 2017

Deldjoo, Y., Frà, Valla M., and Cremonesi, P., *Letting Users Assist What to Watch: An Interactive Query-by-Example Movie Recommendation System*, Proceedings of the 8th Italian Information Retrieval Workshop, IIR 2017, 63–66, Lugano, Switzerland, June 05-07, 2017.

Deldjoo, Y., Elahi M., and Cremonesi P., *Using visual features and latent factors for movie recommendation*, Workshop on New Trends in Content-Based Recommender System (CBRecSys '16) as part of ACM Conference on Recommender Systems 2016, RecSys 2016, Boston, MA, USA, September 16, 2016.

Deldjoo, Y., Elahi, M., Cremonesi P., and Garzotto, F., *Recommending movies based on mise-en-scene design*, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI 2016, 1540-1547, San Jose, CA, USA, May 7-12, 2016, Extended Abstracts.

1.5 Outline of the Thesis

The remainder of the thesis is structured the follows:

- Chapter 2 presents the foundation needed for multimedia processing and recommender system, presents the general framework of content-based recommendation using audio and visual content and positions this thesis with respect to the state-of-the-art in multimedia content analysis for recommendation.
- Chapter 3 investigate the role of visual characteristics of videos for recommendation.
- Chapter 4 extends chapter 3 and presents the role of both audio and visual characteristics of videos by considering the latest descriptors in the field of computer vision and music information retrieval.
- Chapter 5 presents a novel large-scale multifaceted movie trailer dataset (MMTF-14K) to facilitate research on recommendation to facilitate research on video recommendation using multimedia content. It also provides a review common evaluation metrics used to assess the quality of RS in offline settings.
- Chapter 6 Offers the conclusions and future works

CHAPTER 2

Recommendation using Multimedia Content Analysis: State of the Art and Preliminaries

In recent years, we have witnessed an huge amount of multimedia data created by users in various Web channels as well as the social media platforms. As a consequence, it has become harder and harder for users to find interesting new content. Even retrieving a specific (version of an) item sought for sometimes equals looking for a needle in a haystack, especially in the presence of ambiguities or various similar versions of that item. As a countermeasure, recommender systems (RS) that automatically determine content that a user may like have emerged and evolved during the last decade. Since then, a large scale effort of research, mostly algorithmic, has been devoted to improve RS, not least because of initiatives such as the Netflix Prize,[1] driven by commercial interests [191].

As of today, *collaborative filtering* (CF) and *content-based filtering* (CB) are presumably the two most frequently adopted variants of RS today.[2] CF

[1] http://www.netflixprize.com

[2] A RS variant which is not based on this assumption is context-aware or context-based recommendation, which adjust their recommendations based on contextual or situational aspects of the consumption environment.

model assume that a user will appreciate content that is similar to the content other like-minded users prefer, where this similarity is computed either (i) by comparing the user's consumption history with that of different users and suggesting what similar users liked, but the target user has not seen previously (*user-based CF*), or (ii) by taking into account the items in the target user's consumption history, deciding similar items via their co-consumption patterns with other users, and recommending those items with highest similarity, unseen by the target user (*item-based CF*). These similarities are nowadays normally measured not in the original high-dimensional user–item (UI) interaction/consumption space rather in a latent factor space derived from the UI space, by matrix factorization (MF) techniques [24, 190].

In order to alleviate the common issue with CF, such as the cold-start problem when a new user enters the system a or new item is added to the catalog, *content-based filtering* (CB) approaches have been proposed as an alternative or extension to CF, in the latter case to build hybrid systems. CB algorithms base their recommendations on similarities between intrinsic item properties, rather than their co-consumption patterns in users' interactions histories as it is the case with CF. To provide some vivid examples, in the multimedia domain, such item properties can relate to the hues in a picture, the motion trajectories of a movie, the rhythm of a music piece. However, the degree to which CB is used, and in turn the complexity of individual algorithms, strongly varies between different domains. While in the multimedia community, extracting descriptive item features from all kinds of text, audio, image, and video content is a well-established research task, the recommender systems community — for a long time driven by movie recommendation tasks — even has a different interpretation of the term "content". In fact, in RS community, them term content-based is frequently used to refer to (high-level) metadata only, *e.g.,* title, tags, actors, or plot of a movie, disregarding the abundance of information encoded in the actual image and audio signals.

We believe that recommender systems community can strongly benefit from knowledge in multimedia signal processing, built over the last years by the multimedia research community. In this chapter, therefore we follow two main objectives: First, we aim to bridge the gap in perspectives and content analysis advances between the *multimedia* and the *recommender systems* communities. Second, we provide for a general computer science audience a comprehensive review of current state-of-the-art CBRS algorithms that utilize "content" features from a multimedia/signal processing perspective, *i.e.,* features extracted directly from the multimedia content.

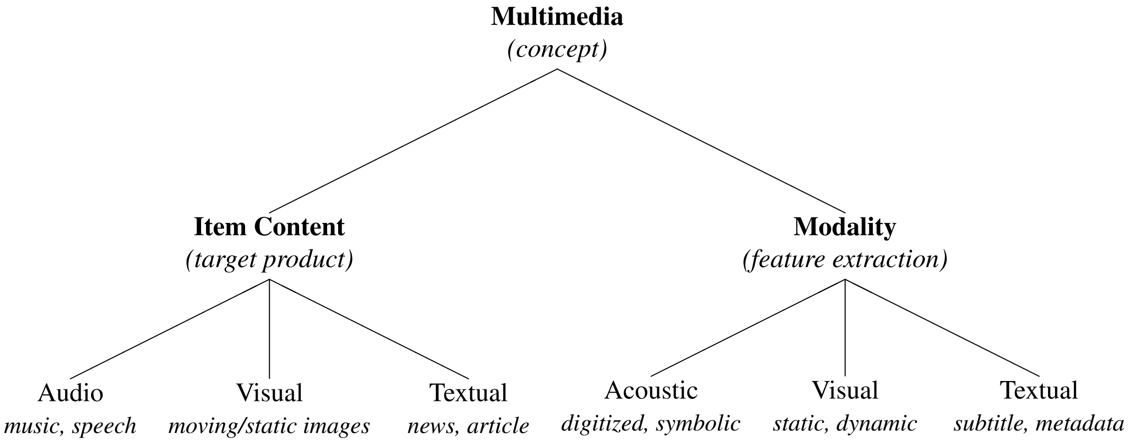


Figure 2.1: *Classification of multimedia item types and modalities for feature extraction*

We also include approaches that integrate metadata together with other forms of multimedia content for the recommendation purpose.

2.1 Foundations of multimedia processing and recommender systems

Multimedia recommender systems, which are the target of this study, consider multimedia material at two levels: (i) as an *item for recommendation* in which each item can be composed of multiple *media contents* such as a video clip, a book or a music piece as shown in the left branch of Figure 2.1(ii) as the *feature extraction modality*. Each media content itself is formally represented in terms of specific *features* which can be of type audio, visual and textual or combination of them as shown in the right branch of Figure 2.1. These features constitute a descriptive representation of the multimedia item upon which the content-based recommender systems are built.

Multimedia features can be classified along different dimensions as we will see later in Section 2.2. One of the distinctive dimensions from a human point of view is the semantic expressiveness of these features. It is common to recognize three levels of expressiveness (aka levels of abstraction), with increasing extent of semantic meaning: *low-level*, *mid-level*, and *high-level* features according to which features can be classified. Low-level features are close to the raw signal (*e.g.,* energy of an audio signal, color in an image, motion in a video, or number of words in a text), high-level features are close to the human perception and interpretation of the signal (*e.g.,* motif in a classical music piece, emotions evoked by a photograph, meaning of a particular video scene, story told by a book author). In the middle of them, mid-level features are more advanced than low-level ones, but further distant from being semantically meaningful as high-level ones. Mid-level features are often expressed as a combination or transformations of low-level features, *e.g.,* by applying human auditory models to the amplitude or frequency representation of an audio signal, or they are inferred from low-level features via machine learning. Some examples from different feature categories are provided in Table 2.1.

Table 2.1: *Categorization of different multimedia features based on their semantic expressiveness.*

High-Level	events, story	structure, mood, message	story, writing style
Mid-Level	objects, people, their interaction	note onsets, rhythm patterns	sentence, term frequency
Low-Level	motion, color, texture, shape	pitch, timbre, loudness	tokens, term frequency
Hierarchy/ Modalities	**Visual**	**Audio**	**Textual**

2.1.1 Multimedia processing

A multimedia item by definition refers to any combination of audio, image, and textual information, as shown in Fig 2.1. Please note that this definition also includes items with a single modality only, which are sometimes not considered multimedia items though. The main categories of audio items are music and speech. The main categories for the visual media type include static images (like a picture) and moving images (like a movie). Textual information can be pure text like a book without any images, text with visual information such as a web page containing texts and several images, and spoken textual information recorded orally, such as speech. For simplicity of the illustration, in Fig 2.1 we only show the interconnection between neighboring modalities, whereas in reality, for instance, speech should reside in the intersecting point between textual and audio information. Also video typically contains moving images as well as audio, sometime also textual information (e.g., subtitles). In the following, we will introduce the fundamental concepts in Audio Content Analysis (ACA), Image Content Analysis (ICA), and Video Content Analysis (VCA).

Table 2.2: *Physical and psychological dimensions of music. Courtesy of [101].*

physical characteristics of sound waves	psychological perception of music
frequency [Hz]	pitch
amplitude [dB]	loudness
complexity	timbre

Audio Content Analysis:

Fundamentally, on the physical level, sound is created as a result of propagation of energy from a source into a medium such as air (or water). The vibrating energy acts on the surrounding air molecules, compressing the ones that are in contact with the vibrating material. These excited molecules in turn, compress the molecules adjoining them, and this progressive patterns propagates through the air until it arrived the ear drum. A *sound wave* is therefore generated as the result of successive compression and expansion of air molecules, which can eventually excite the human auditory system, giving rise to the perception of sound [101].

The branch of science that is concerned with sounds (creation of sound, its transmission and perception) is known by *acoustics*. *Musical acoustics* is a specific sub-domain of the acoustics field whose focus is on the mechanisms of sound production by musical instruments, impacts of reproduction processes on musical sounds, and human perception of sound as music [101, 138]. The sound produced at the source side and what is how it is perceived by the human auditory system are not the same. [101] presents an interesting study by comparing the physical characteristics of a sound wave and the perception of those characteristics focusing on the psychological dimensions. The main low-level properties representing sound and music by contrasting their physical characteristics and their psychological perceptions is shown in Table 2.2.

- **Frequency and pitch**: The primary parameter identified with a sound wave is the rate at which it alternates: the *frequency* measured in Hertz (Hz) or cycles per second. For instance, a sound wave vibrating at 40 cycles per second has a frequency of 40 Hz. The frequency of the sound wave gives rise to the aural sensation of *pitch*. Human can discern the pitch range lying in the frequency range from 20 Hz to 20000 Hz. By the age of 30 or 40, as the result of a phenomenon known as *presbycusis* (or hearing loss associated with age), the sensitivity of the human auditory system to high-pitch ranges decreases. Nevertheless this hearing loss does not appreciably affects the music listening experience since music is typically played in the frequency range 20 Hz to 4000 Hz [101, 199].

There is a positive mathematical correlation between the frequency f of the audio signal at the sound source and its perceived pitch in the sense that a high frequency tone results in perception of it as high-pitched while a low frequency signal leads to a low-pitched perceived sound. In high frequency ranges, this relationship is non-linear where two pitches with the same perceived pitch distance have a larger frequency distance than at lower frequencies [199]. This non-linearity is caused by the frequency resolution of the human cochlea.

The *Mel scale* model is one of the most common models for mapping the actual frequency to its perceived pitch given by Equation 2.1, where f denotes the frequency in Hertz.

$$m(f) = 1127 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \; [Mel] \tag{2.1}$$

According to the Mel scale, the relationship between Hertz and Mel is approximately linear below 1000 Hz and logarithmic above. The scale's reference point is 1000, which means that $1000\,Hz = 1000\,Mel$ [180, 332].

Motivated by the field of physiology, in pitch processing, there exists a concept named *critical bandwidth* or just *critical bands* of the human hearing system, which refers to the extent of frequencies that evoke a comparative sensation in the auditory system [101]. It is based on the fact that the human auditory system responds to the incoming frequencies reaching it selectively. For example, if the incoming sound reaching the ear drum contains several frequencies in close selectively with each other, they are perceived as one pitch. The *Bark scale* is constructed based on this effect; it quantizes music signals into critical bands whose range corresponds to that of human auditory perception. In other words, in each band, the lowest and the highest frequency correspond to one pitch interval [179]. The function to convert Hertz to Bark is given in Equation 2.2.

$$b(f) = \frac{26.81 \cdot f}{1960 + f} - 0.53\,[Bark] \tag{2.2}$$

For example, the frequency range [1250, 1460] Hz corresponds to [10, 11] on the Bark range while the frequency range [4370, 5230] Hz corresponds to [18, 19] on the Bark scale [180]. These models serve as approximation models in many practical audio applications [101, 180, 199].

- **Amplitude and loudness**: The second relevant parameter to sound and music is the *amplitude*. The amplitude of a wave is the maximum displacement of the points in the wave. The greater is the amplitude, the more energy it transmits into the medium. The amplitude of the signal (sound wave) creates the aural sensation of *loudness*. Sound waves with higher energy (*i.e.,* larger amplitude) are heard louder than sound waves with less energy, at the same frequency [101]. The physical loudness of a sound is known as *intensity* measured in decibel (dB) according to Equation 2.3, in which i is the energy of the sound and i_0 is the energy of the faintest discernible sound (human hearing threshold), *i.e.,* 10^{-12} watts/m^2 (at frequency 100 Hz) or 10^{-16} watts/m^2 (at frequency 3000 Hz).

$$i_{dB} = 10 \ \cdot \ \log_{10}\left(\frac{i}{i_0}\right) \ [dB] \tag{2.3}$$

  This means if we give the faintest describable sound a value of $i_0 = 1$, the loudness 0 dB represents the hearing threshold since $10 \cdot \log \frac{1}{1} = 0$. Doubling the sound energy would yield the loudness equal to 3 dB since $10 \cdot \log_{10} \frac{2}{1} \approx 10 \cdot 0.3 = 3$. To give some real-world examples, a whisper at a distance of 1 meter has a loudness about 10 to 15 dB whereas a normal conversation at the same distance is approximately 65 dB. The environment sound in a big city or house is about 40 dB. The human auditory system allows discrimination of sound waves with differences in their loudness of approximately 1 dB for frequencies between 200 Hz and 4000 Hz.

- **Timbre and Complexity**: The third related parameter is the *complexity* of the sound. Sine waves with a single frequency represent a 'tone' in music. However, most sounds do not contain a single tone, rather an arrangement of many, occurring simultaneously. Such sounds can be regarded as complex tones, and the interplay between the constituent frequencies yields a characteristic of sound, known as *timbre*. For instance, a beep sound or the sound produced by a tuning fork are considered single tones, while a friend's speech or piano play and regular sounds in nature have complex descriptions. The composing frequencies (harmonics) of a complex sound signal can be decomposed by Fourier analysis [101, 150].

Image Content Analysis:

A digital image is characterized via a two-dimensional function $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ where x and y are the *spatial* (plane) *coordinates*. In practice, the spatial coordinates x and y are typically restricted to a rectangular shape thereby $f(x, y) : [a, b] \times [c, d] \rightarrow \mathbb{R}$. The value of the function f at any pair coordinates (x_0, y_0) denoted by $f(x_0, y_0)$ indicate the *brightness* of *intensity* at that pixel. Three types of images based on the value of $f(x, y)$ can be identified: (i) *monochrome image* (aka. black & white image), in which the output is limited to two colors, black (0) and white (1), (ii) *gray-scale image*, in which $f(x, y)$ represents shades of gray color, varying from 0 (black) at the weakest intensity to 2^n (white) at the strongest where for limitation of HVS, n is usually set to 8, and (iii) color image in which the value of $f(x, y)$ at each pixel is given by a 3-tuple (R, G, B) denoting the amount of redness, greenness, and blueness in that pixel.

Color can be comprehended as an electromagnetic (EM) wave hitting an object and reflecting toward the observer's eye where in this process some wavelengths are absorbed and some reach the eyes, giving the perception of different colors. Visible colors lie within the wavelength range 380–780 nm known as the *visible spectrum*, where each wavelength characterizes an specific color; at the extreme short and long wavelengths reside, respectively, violet and red, while between them there are shades of green and yellow. The perception of color happens at retina due to the existence of three types of color photo-receptor cells namely *cones* which respond to wavelengths near the blue, red, and green lights as shown in Fig 2.2. Another of type of photo-receptor cell known as *rods* in the retina is responsible for vision during night and is effective only at extremely low light levels. Inspired by the human visual system (HVS), colors can be represented by three stimuli according to the way they are perceived, known as *tristimulus representation*. Developed formally by Thomas Young and Hermann von Helmholtz [295], the *tristimulus theory* states that all possible colors humans perceive can be represented in a 3D linear space defined by the *colorimetric* equation, given by Equation 2.4, in which $[A_1, A_2, A_3]$ define the *base colors* (or primaries) and $[c_1, c_2, c_3]$ are the associated *weights* with each color[3].

$$\mathbf{C} = c_1\, A_1 + c_2\, A_2 + c_3\, A_3 \tag{2.4}$$

The set of all colors form a vector space are called *color space* or *color model*. The three components of a color can be defined in many different

[3] The equality does not mean that the algebraic summation in the right side gives a numerical value C that can be used to represent or re-create the color. The symbol C is not a value or a color representation, but the equation expresses the idea that three stimuli combined by superposition of lights re-create the perception of the color.

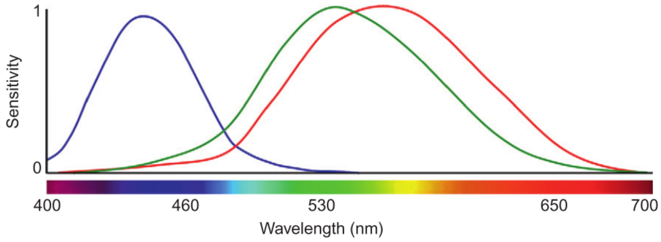


Figure 2.2: *Visible spectrum and tristimulus response curves. Courtesy of [236].*

ways leading to various color spaces.

For example if we define the base colors in Equation 2.4 based on the human visual system (HVS), the primary colors will need to be set as the response of the cone receptors in the human eye: red, green, and blue. This model is formally known as the *CIE RGB model* given by Equation 2.5.

$$\mathbf{C} = r\,R + g\,G + b\,B \tag{2.5}$$

The CIE RGB color model considers how colors are perceived by the human eye and was developed on such a premise; however it has several undesirable properties as discussed in [236]. For example, one of this undesirables effects is that all three base colors contribute to the definition of brightness (or illumination) of a color. Due to high sensitivity of the HVS on perception of brightness, there is a need for a color model to concentrate on the brightness as a single component. The CIE XYZ model, given in Equation 2.6, was developed as a universal reference system to combat these unwanted properties.

$$\mathbf{C} = x\,X + y\,Y + z\,Z \tag{2.6}$$

The color components in the XYZ color model can be calculated directly from the components of the CIE RGB model by a linear transformation given by Equation 2.7, where M is a predefined 3×3 non-singular matrix, defining the mapping from the CIE XYZ to the CIE RGB color model.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = M \begin{pmatrix} r \\ g \\ b \end{pmatrix} \tag{2.7}$$

Generally there is no technique to determine the optimum color space model for all multimedia applications. The choice of a color model heavily depends on properties of the model and the design characteristics of

Table 2.3: *Classification of color models. Courtesy of [250]*

type	model	application
user-centric	HSI, HSI, HLS	human color perception, computer graphics
device-centric	RGB, YIQ, YCC (non-uniform space)	color difference evaluation, color management system
	LAB, LUV (uniform space)	coding, color TV, storage

the application. The most popular color models and their applications are summarized in Table 2.3.

Computational features extracted from images share the common goal to convert the image (or segmentations of the image) into a representation that better describes the image's main properties. Feature extraction is the critical step in many computer vision tasks since it enables the transition from a qualitative (pictorial) representation into a quantitative (non-pictorial) representation which can be exploited in pattern recognition, classification, and recommendation tasks. Two types of features are extractable from images and videos, depending on the application at hand: (1) static features (aka spatial features) and (2) dynamic features (aka temporal features), where (1) can be used for both ICA and VCA and (2) is specific to VCA.

Static features are extracted from an image or single frame of a video, often the *key frame* (see Section 2.1.1). For videos, these features cover the visual characteristics of the video to some extent. Main examples of static features are *color*, *texture* described in the following:

- **Color:** Color is a shared property between human perception and computer vision. It is perhaps the most expressive and extensively utilized feature in image and video retrieval systems. Color has been used in various fields ranging from purely scientific, to abstract art, and applied applications [219]. Using color as a low-level descriptor in recommender systems of pictorial objects such as images or videos is crucial since color does not only allow discrimination of objects by adding information to them, but it also adds beauty to objects. There exists a strong emotional and psychological impact of color on humans which the system can explore in various ways and applications (*e.g.,* recommendation of paintings in an art museum). The studies on emotional and psychological influence of color originally dates back to 1840 in the famous book by Goethe named "Farbenlehre" [200, 324]. From a processing point of view, colors are characterized by a specific color space as described in the previous section.

In image retrieval, the color spaces RGB, LAB, LUV, HSV (HSL), YCrCb and the hue-min-max-difference (HMMD) are widely used since they are known to be closer to human perception. Standard color descriptors include but not limited to color-covariance matrix, color histogram, color moments, and color coherence vector [211]. Standards like MPEG-7 have taken the effort to formally define content-based descriptors that can be leveraged to build search, retrieval and recommender systems in order to effectively and efficiently identify, browse, or retrieve images. The MPEG-7 color and texture visual descriptors are summarized in Table 2.4 [64, 200, 211, 219].Two software implementation of for extracting and matching MPEG-7 visual descriptions from associated visual content of images can be found in [7, 8].

Table 2.4: *Visual descriptors for color and texture as defined by MPEG-7 standard. Offical homepage can be found in [9].*

MPEG-7	features names
color	dominant color (DC), color structure (CS), scalable color (SC), color layout (CL)
texture	homogeneous texture (HT), edge histogram (EH)

- **Texture:** Texture is the second powerful descriptor of image. Texture however has not been employed in image and video retrieval systems so often perhaps because it is not well-defined. Texture is an intuitive concept. Although we humans are able to discern texture when we see it, for example spots on leopards *v.s.* stripes on tigers, it may be difficult to characterize a precise formulation for texture. Nevertheless, there is a agreement on two main properties of texture: (i) there exists a significant changes in intensity levels of the nearby pixels within a texture which is equivalent to say at finer resolutions, texture has a non-homogeneous structure however, (ii) texture is a homogeneous property at some spatial scale larger than the resolution of the image [324]. In the literature, three main approaches for analyzing textures re suggested [200, 211]:
  - *Statistical approach:* The characteristics of texture is described by a set of features such as contrast, correlation, and entropy. Detailed information can be found in [140].
  - *Stochastic approach:* It is assumed that a stochastic process governed by some parameters leads to texture. The estimated parameters of the models serve as the features for texture classification

and segmentation. Detailed information can be found in [137].

- *Structural approach:* This approach is applicable to certain types of textures, which are viewed as two-dimensional patterns consisting of a set of primitives or subpatterns arranged according to specific placement rules. Some examples of such textures include tilings of the plane, cellular structures such as tissue samples, and a picture of a brick wall. Detailed information can be found in [137, 140].

A common requirement for feature extraction and representation techniques in ICA is that the features are robust to changes in rotation, scaling, and translation, collectively known as RST. Being *RST-invariant* is an advantage in some applications because it can ensure that a machine vision to capture the same low-level/semantic information from an image independent of what size, position within the image, and angle its appear at [221].

Video Content Analysis:

A video is a sequence of images extended over time, mathematically defined as $f(x, y, t)$ where $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ for a gray-scale video and $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ for a color video using a 3-dimensional color space. Each image constitutes the basic unit in a video, known as *frame*. Consecutive video frames mostly contain a high degree of similarity in their visual content and give the user the illusion of motion.

Because consecutive video frames are visually highly correlated, considering all such frames for VCA is computationally inefficient due to high amount of redundancy in the data. Therefore, it is common to split the video into semantically meaningful segments known as *video scenes*, where each scene can contain several shots, in which a *shot* is a single camera action. For instance, a conversation scene between three people can contain several shots capturing each speaker when he/she speaks, but they all belong to one semantic unit, the conversation scene. It thus makes sense to capture information from the video scene that can tell us what is happening in the scene. This kind of information are also referred to as *semantic features*, as illustrated in Fig 2.3.

Frames around the shot boundary have a significant dissimilarity in their visual content, while frames within a shot are highly similar. It is thus common to select a single frame (*e.g.,* the middle frame) from each shot and use that frame for feature extraction. This frame is called the *key frame*. Fig 2.3 illustrates the hierarchical representation of a video. Two types of low-level features are then extractable from videos, depending on the application in hand: (1) static features (aka spatial features) and (2) dynamic features (aka temporal features).

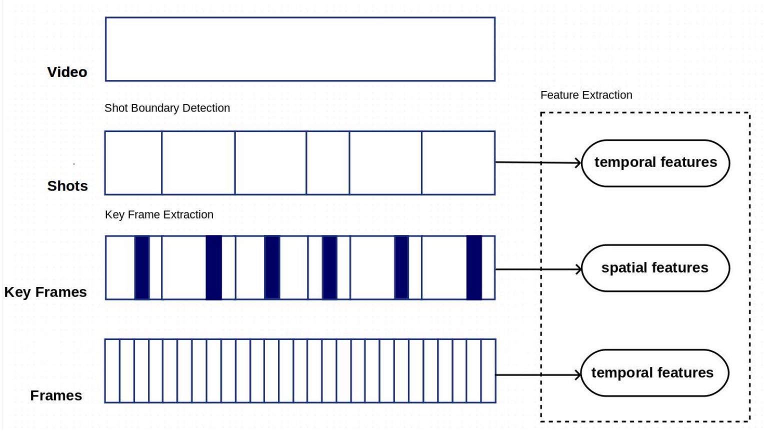


Figure 2.3: *Temporal segmentation in a video. Courtesy of [90].*

In addition to static and dynamic features extracted from the visual channel of a video, many videos also contain audio channels and/or text (e.g., subtitles), to which dedicated feature extraction techniques can be applied to obtain a more comprehensive description of the video.

Dynamic features: Motion is an essential aspect of video sequences, distinguishing them from images. Motion can be due to camera movement or object movement. The ability to estimate motion is important for many tasks (retrieval, classification, or recommendation). For example, certain genres are known to be made with fast pace (e.g., Action movies) while others are typically shot slow-paced (e.g., Drama). Motion estimation is the process of analyzing successive frames of a video in order to identify objects that are moving. The latter is usually realized by estimating a series of two-dimensional vectors called motion vectors (MV), which encode the length and direction of motion. According to [221], we can identify between different types of motion estimation methods:

1. Still camera, single moving object, constant background.
2. Still camera, several moving objects, constant background.
3. Still camera, single moving object, moving (cluttered) background.
4. Still camera, several moving objects, moving (cluttered) background.
5. Moving camera, fixed object(s), constant background.

6. Moving camera, several moving objects, moving (cluttered) background.

The first and the second problem are addressed/solved by most motion estimation algorithms in use today. When the background is complex (e.g., wavering tree branches), the task becomes more complicated though, and a significant amount of ongoing work in VCA deals with those scenarios. The fifth problem is a separate problem and entails usage of camera movement detection methods. The last problem is a sophisticated problem and beyond the reach of contemporary VCA systems. Further information about motion and its estimation can be found in [155, 221].

Recommender systems

Recommender systems (RS) are arguably one of the most essential accomplishments of machine learning (ML) in the last decade. They refer to a set of software tools and algorithms which provide suggestions to users for items that are most likely interesting for them [23, 268]. These suggestions assist the users in decision-making about what product to purchase or which music to listen to. In the multimedia domain, the items are of types audio (*e.g.,* a piece of music), visual (*e.g.,* a movie on Netflix or painting in an art gallery), or textual (*e.g.,* a news on the web site), as shown in Fig 2.1. Effective recommendation often requires the knowledge about of the properties of specific items at hand, the graphical user interface, and the core recommendation technique used to produce the recommendations [268]. The previous components are all customized to provide valuable and effective recommendations for the particular type of item. Recommendation are appreciated by users when they are *personalized* to their tastes and interests[4]. Therefore in order to be able to provide such functionality, such system require to collect information from users regarding their preferences either expressed explicitly, *e.g.,* as ratings for products, or inferred implicitly by interpreting their actions on the items. Formally, let $u \in U$ be a user from the set of all users, $i \in I$ be an item from the set of all items, and $r \in R$ be a rating associated with a user-item pair, i.e., an ordered set of non-negative numbers, then $g : U \times I \to R$ is the *utility function* which measures the usefulness of item i for user u. For each user $u \in U$, we aim to choose an item $i \in I$ that maximizes the utility given by Equation 2.8.

$$\forall u \in U, \quad i_u = \arg\max_{i \in I} g(u, i) \tag{2.8}$$

[4] Example of non-personalized RSs in the movie domain include *top-popular* movies, *top-rated* movies or *random* movie recommendations.

We refer the user for which recommendations are provided as *active user* or *target user*, both terms used interchangeably. Given this formal definition, the goal of a RS is to learn the utility function g based on given rating data, in a way that g is capable of accurately predicting a rating for *unknown* user-item pairs and thus simulating the user behavior. Three main strategies exists for this purpose: collaborative filtering (CF), content-based filtering (CB), and hybrid models.

- **Collaborative filtering (CF)** exploits the *interactions* between users and items expressed through their preference to make recommendations. In other words, CF models leverage the fact that the specified preference (*e.g.,* rating scores) are often highly correlated across users and items, therefore the unspecified ratings (to be predicted by the RS) can be imputed by considering the *inter-item* correlations (*item-based* CF model) or *inter-user* correlations (*user-based* CF model). A few CF models utilize both types of correlations. CF-based models are classified in two types: (1) memory-based and (2) model-based. The distinction is that in (1) the ratings of few similar users or items are aggregated (typically using the K-nearest neighbor (KNN) method) to predict an unknown rating, where similarity is characterized using a similarity metric (usually the Pearson correlation or the cosine similarity), whereas in (2) the recommendation is enabled based on prediction models that have been trained using the user-item matrix, in whole or in part. Such a trained prediction model is provided as the input for making recommendations for individual users [16, 23, 111, 297]. The measured prediction models represent users and items in a joint latent space which is usually of much lower dimensionality than the original rating space, making recommendations more efficient to compute.

  One of the standard measures to compute correlation between users in a user-based CF model (by removing the effects of mean and variance) is the Pearson correlation similarity defined by Equation 2.9, in which s_{uv} defined the similarity score between the ratings expressed by users u and v. In order to provide an example for rating prediction for an unknown user-item pair $\hat{\mathbf{r}}_{ui}$ in a simple user-based CF system, consider Equation 2.10, where $N_i(u)$ denotes the users who have rated item i and are most similar to user u, since only their ratings can be used in the prediction of $\hat{\mathbf{r}}_{ui}$ [25, 235].

$$s_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt[2]{\sum_{i \in I_{uv}} (r_{ui} - \overline{r}_u)^2 \sum_{i \in I_{uv}} (r_{vi} - \overline{r}_u)^2}} \tag{2.9}$$

$$\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{i \in N_i(u)} s_{uv}} \sum_{i \in N_i(u)} s_{uv} \, \mathbf{r}_{vi} \tag{2.10}$$

Since CF methods can somehow leverage the shared knowledge between community of users (through their behaviors), they are known to be able to provide high-quality recommendations when sufficient information about users' interactions are available. The fundamental challenge with CF strategies is that the underlying user-rating matrix is typically very sparse a problem known as the *sparsity problem* which can introduce negative impact on the effectiveness of CF systems. Also when a new user or item has newly entered the system, CF approaches are unable to generate useful recommendations due to the lack of sufficient previous interactions [23, 157].

- **Content-based (CB)** recommenders use content descriptors of items typically describe in the form of feature vectors, to make recommendations. CB models combine the preference indications of the target user and content information available about the items in order to build a user model (aka user profile) and compare it with the descriptive information of the content (aka item profile). These models have the advantage that they do not require the rating of other users, therefore as long as the target user's *own* preference is available, they can make recommendations. The other users' ratings hence do not affect CB systems which is beneficial in cold-start scenarios [19].

  At the most fundamental level, CB models require two sources of information to compute recommendations: (1) *item profile*, which is the content-based attributes of items, for example color in an image or instrumentation of a piece of music, and (2) *user profile*, which is computed from users' feedbacks on items, implicitly (clicking on item) or explicitly (expressed by rating). The CB user profile is a feature vector in the vector space spanned by the item features determining the amount of user interest in each of the items/features. A CBRS compares the user profile of the active user with the item profile of various items and recommend the item with the closest match of the profiles, usually disregarding the items the user is already aware of.

Because of the relevance of this thesis (and this chapter) to CB models, in the following we define some of the variations of such systems. The most widely used algorithm in CB systems is the item-based K-*nearest neighbor (KNN)*, in which the unknown preference score (rating) $\hat{\mathbf{r}}_{ui}$ for user u and item i is computed as an aggregate of the ratings of u for similar items [16, 235]. Standard aggregation functions include (1) the average and, (2) the weighted average, given by Equations 2.11 and 2.12, respectively,

$$\hat{\mathbf{r}}_{ui} = \frac{1}{K} \sum_{j \in N_u(i)} \mathbf{r}_{ui} \tag{2.11}$$

$$\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{j \in N_u(i)} s_{ij}} \sum_{j \in N_u(i)} s_{ij} \, \mathbf{r}_{ui} \tag{2.12}$$

where $N_u(i)$ denotes the items rated by user u most similar to item i (*i.e., i*'s neighbors) and s_{ij} is the similarity score between items i and j. The difference between a CF item-based and CB item-based model lies in the computation of s_{ij}, the former computes the similarity between rating scores by co-rating users and the latter using the content-based notion of similarity between item descriptions. It is also possible to use s_{ij}^{α} rather than s_{ij} where $\alpha > 0$ is an amplification factor whose role is to assign greater importance to the neighbors that are the closer to i [50, 235].

Another common approach to CB recommendation is based on *regression models*. These models have the merit that they can be used for various types of preference scores, including binary, numerical, or interval scales. [19] provides a comparison of different model of regression family that can be used for the recommendation problem depending on the application at hand [19]. A CB model based on linear regression is given by Equation 2.13,

$$\widetilde{\mathbf{r}}_{ui} = \alpha + \mathbf{p}_u \mathbf{f}_i^t \tag{2.13}$$

in which α is a scalar bias term, $\mathbf{p}_u$ is the $1 \times n_F$ regression coefficient vector, $\mathbf{f}_i$ is the $1 \times n_F$ feature vector of item i and $\widetilde{\mathbf{r}}_{ui}$ is the estimated rating for user u on item i. Here, $\mathbf{p}_u$ can be viewed as the *user profile* for user u, a weight vector measuring u's taste on each of the feature vector components. The user profile $\mathbf{p}_u$ can be estimated by the ridge regression optimization model [249], given by Equation 2.14, where $\|.\|_2^2$ is the ℓ_2-norm and the constant $\lambda > 0$ is the regularization parameter.

$$\min_{\mathbf{p}_u} \frac{1}{2}\|\mathbf{r}_{ui} - \mathbf{p}_u \mathbf{f}_i^t\|_2^2 + \lambda\|\mathbf{p}_u\|_2^2 \tag{2.14}$$

The optimization model in Equation 2.14 has an closed-form solution which learns the user profile $\widetilde{\mathbf{p}}_u$ in the train phase and uses it predict the unknown rating in the test phase [19]. It is also possible to use L1-regularization by replacing the regularization term $\|\mathbf{p}_u\|_2^2$ with $\|\mathbf{p}_u\|_1$. This form of regularization does not have a closed-form algebraic solution and gradient. descent methods must be used. Using L1-regularization is known as Lasso [122], which can be used in a dual role for feature selection since it will activate only a few coefficients, resulting in a sparse coefficient vector or *sparse solution* [19].

CBRS have the peculiarity which is they often require to gather item descriptions coming from different types of unstructured data and convert them into standardized representation. To this end, some additional steps are often required, including data pre-processing, feature extraction, and feature post-processing, which are intrinsic part of CBRS. In this chapter to the relevance of the topic to CBRS, we will describe these additional steps in more detail with the focus on multimedia content.

- **Hybrid models** combine different RS approaches, using different sources of information, each having their own strengths and weaknesses. For instance, while CF methods utilize the ratings of a *community* of users, CB methods use the ratings of a the active user's *own* preference together with item descriptions (features) in order to make recommendations. Hybrid RS combine at least two recommendation models and exploit the the algorithmic power and knowledge available in each of recommendation models in order to make robust inferences. [55] present different classes of hybridization for RS including: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level hybrids. In RS literature, sometimes terminologies like "CF with side information" are invented to describe the hybridization from a RS point of view. In this chapter, we however resort to the systematic definition of hybrid systems as presented in [55]. We should also remind that from a systematic point of view we consider a CBRS using two types of MM features (*e.g.,* color and texture) a hybrid recommender while in RS terminology the former can be considered a pure CB system. A good presentation of hybrid RS can be found in [21, 268]. We describe some of the important techniques in the following.

In recent years, a new family of RS has emerged known by *cold-start recommender systems*, used primarily to recommend new items. These recommendation techniques typically have a hybrid regression-based nature and employ some parameter learning during the train phase. One of such methods is the *attribute-to-feature mapping* (AFM), defined by Equation 2.15 [123]

$$\widetilde{\mathbf{r}}_{ui} = \mathbf{p}_u \, A \, \mathbf{f}_i^t \tag{2.15}$$

where $\mathbf{f}_i$ is the $1 \times n_F$ feature vector of item i, $\mathbf{p}_u$ is the $1 \times n_F$ regression coefficient vector, and A is the $n_F \times n_F$ matrix of regression coefficients. In cold-start settings, the AFM method the factorization terms $R = PQ^t$, *i.e,* maps the preferences of users and items into the latent space . Next, another mapping $R = PQ^t = PAF^t$ is learned between item latent representation and their feature representation, which eventually defines the recommendation model of AFM. Bayesian personalized ranking (BPR) is used to learn the factorizations [123, 264].

An alternative to AFM is a technique known as the *regression-based latent-factor model* (RLFM) [18], defined by Equation 2.16

$$\widetilde{\mathbf{r}}_{ui} = \alpha + \mathbf{b} \, \mathbf{f}_i^t + \mathbf{p}_u \, A \, \mathbf{f}_i^t \tag{2.16}$$

The main difference between the AFM and RLFM model is in introduction of new term $\mathbf{b} \, \mathbf{f}_i^t$ in RLFM which represents users' global preferences toward item i, also known as *feature-based item bias.* In addition, while in AFM the model parameters are leaned in two separate steps, RLFM learns them simultaneously in a single step [18, 114, 123].

The last very popular hybridization technique is the Factorization machines (FM) [262] which is a model taking advantages of both support vector machines (SVM) and factorization models (FM). The innovative idea behind FM is to transform user-item interactions and the features associated with users and items into a compact real-valued feature vector and leveraging a strong SVM-like learning algorithm in a regression framework. FM can be formally defined by Equation 2.17 (bias terms are omitted for simplicity),

$$\widetilde{\mathbf{r}}_{ui} = \sum_{c=1}^{k} \sum_{c'=c+1}^{k} x_{ui,c} \, x_{ui,c'} \mathbf{v}_c \, \mathbf{v}_{c'}^t \tag{2.17}$$

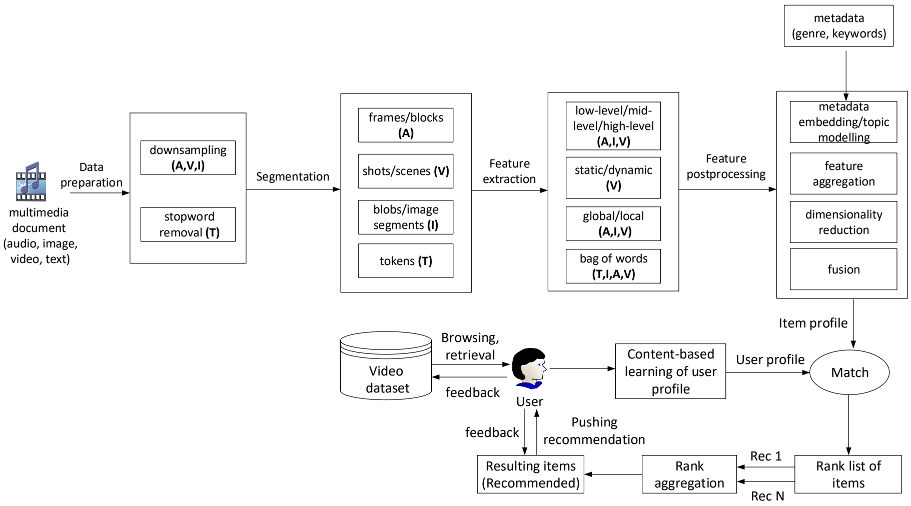


Figure 2.4: *Generic framework for a recommender system using content-based multimedia analysis*

where k is the length of the feature vector, $x_{ui,c}$ is the value of feature c in feature vector x_{ui}, and $\mathbf{v}_c$ is the vector of latent factors representing feature c. FM allows to model higher-order interactions in a way different from classical approaches (such as tensor factorization, *e.g.,* [297]). In particular, since FM can capture feature interactions, they are known to be powerful for handling heterogeneous information sources and can thus represent a promising approach for building content-based multimedia RS when multiple modalities are used (*e.g.,* audio, visual, and/or textual). Further information can be found in [114, 262, 297].

2.1.2 General framework

The main processing stages involved in a content-based multimedia RS is illustrated in Fig 2.4. In the following, we describe each of the stages with differentiation between different media contents.

Data preparation

To apply existing data mining techniques on multimedia data, the first step usually is comprised of transforming data into suitable format for mining [43]. For some type of multimedia data, this step is particularly fundamental. For instance, the video data as a whole is very large data composed of many similar frames compressed together in time. The general aim of data preparation step is to remove noise and redundancies from the multimedia, for example the latter because a naturally-looking video is composed of many highly similar frames. Typical tasks in the data preparation step include data cleaning, normalization, transformation, and feature selection. In the following, we discuss common approaches, separately for each modality.

(i) *Audio preparation:* In the audio domain, feature extraction is not performed on the continuous, analog signal produced by real-world instruments, instead the continuous audio signal $x(t)$ needs to be discretized in time (known by *sampling*) and in amplitude (known by *quantization*):

  (a) Sampling: The discretization in time is done by sampling the signal at uniform intervals of T_s times per second between sampling points. In picking the correct value for T_s, we need to take into consideration the sampling theorem stated as follows: "the sampled signal can be reconstructed without loss of information if $f_s > 2 \cdot f_{max}$, where f_{max} is the the highest frequency in the signal" [180, 199].

  (b) Quantization: Quantization refers mapping each amplitude of the signal (which is sampled in time) to a pre-defined scale of allowed amplitude values, usually represented with a binary code of power of 2, where the codeword length is given by $n = \log_2 M$, M being the number of levels and n the *resolution of the analog-digital conversion.* A common practice in MIR is to use an uncompressed file format like WAV (instead of a compressed format like MP3) with sampling rate of 44.1 kHz, *i.e.,* samples per second, and a resolution of 16 bits per channel.

(ii) *Image preparation:* The image preparation step refers to preparing the digital images with the right spatial resolution and gray-level (intensity) resolution defined according to the following:

  (a) Spatial resolution: According to [130], "spatial resolution is the smallest perceivable detail in an image". It quantifies the amount of pixels per certain fixed physical size (usually inch) and can be used to compare quality (clarity) of two images. The greater the spatial resolution, the higher the density of pixels in an image [6, 130, 221].

(b) Gray-level resolution: "Gray-level (intensity) resolution refers to the smallest change in intensity level that the human visual system (HVS) can discern" [221]. It is known today that adoption of 256 intensity levels (8 bits per pixel) is practically a good choice. It is worthwhile to note that some image applications require higher gray-level resolution, thus formats such as 12-bit RAW or 16-bit tagged image file format (TIFF) can be used for this purpose.

(iii) *Video preparation:* Data preprocessing for videos commonly means to prepare the video at a specific frame rate (see section 2.1.1). The value of frame rate affects mostly dynamic features since motion features are directly computed from the succession of frames. Static features are less affected by this choice since they are computed on a single frame. Nevertheless, the evolution of static features over the span of time may have an influence on the quality of video recommendation. Practical choices for decomposing videos into frame are 5 or 10 frames/second; this is while many practical papers coping with large scale videos [14] prefer to choose the frame rate as low as 1 frame per second (fps) and avoid the temporal segmentation step in order to reduce the computational time and storage requirements in the feature extraction step. In this regard, in choosing the right frame rate, it is worthwhile to pay attention to the specific video document in hand; for example, since movie trailers are specifically designed to serve as an exciting media, they are usually made with lots of abrupt cuts, while full movies are made with a eye-looking natural pace; as the results the extraction of motion features (and thus setting up a right frame rate) may be less vital for movie trailers and more necessary for full movies.

Segmentation:

The goal of segmentation is to split the multimedia item (audio, video, or image) into smaller disjoint structural units each having similar semantic content [155].

(i) *Audio segmentation:* In the audio domain, after having converted the signal from analog to digital, to compute meaningful acoustic features, consecutive digital samples are aggregated into large chunks long enough to be perceivable by the human ear. Segmentation approaches commonly operate at the *frame-level* or the *block-level*, the latter sometimes referred to as *segment-level*.

(a) *Frame-level segmentation*: An audio frame is usually comprised of 256 to 8192 consecutive samples. The human ear has a sensitivity resolution of tens of milliseconds (ms), which needs to be considered when deciding on the frame size in relation to the sampling frequency [179]. For instance, a frame size of 1024 for a signal sampled at 44.1 kHz yields frames that cover about 23 ms of the signal. The resulting frame representation can be used as input to feature extractors that operate in the time domain. For features that are computed in the frequency domain, we need to perform *windowing* before computing the transform. This operation refers to applying a sliding window function to each frame. Mel-Frequency cepstral coefficients (MFCCs) have proven to be one of the most useful *frame-level features* for many audio and music processing tasks [115, 213].

(b) *Block-level segmentation*: Blocks are created by considering larger segments of a piece, typically a few seconds. Due to this property, the temporal essence of audio can be better captured by this segmentation scheme. In [289], six different *block-level features* (BLFs) are presented and a method to fuse all the blocks together. BLFs have shown considerable performances in the MIREX challenges. [5]

(ii) *Image segmentation*: Also known as spatial segmentation, image segmentation is a fundamental problem in image processing and computer vision since it marks the transition between low-level image processing and image analysis. Image segmentation aims to decompose an image into several disjoint partitions where each partition represents a semantically meaningful part of the image. Once an image has been segmented, the resulting individual partitions (or objects) can be described, represented, analyzed, and classified [221].

For a recommender system practitioner, image segmentation can be useful in an application where parts of the image needs to be classified into meaningful descriptions to serve as a semantic content for the system. For example, in order to determine if an image is representing an indoor or outdoor scene, one approach can be to segment the upper portion of the image and calculate the color histogram distribution on the upper portion of the image based on the segmentation results. If the color has a tendency towards white and blue colors, it can be

[5] Annual Music Information Retrieval eXchange (MIREX). More information is available at: http://www.music-ir.org

an indication of sky in the scene so the image can be considered as outdoor, or indoor otherwise. Image segmentation can be also useful if it is desired to extract computational features from some particular regions in the image (*local feature extraction*), instead of the entire image (*global feature extraction*). However, many of today’s image RS employ a holistic global technique [245].

(iii) *Video segmentation*: Generally, the goal of video segmentation is to partition the video into segments that are *homogeneous* in some feature space. Depending on whether the partitioning is effected on individual images (video frames), along the time domain, or both, the segmentation is known as *spatial segmentation*, *temporal segmentation*, or *spatio-temporal segmentation*, respectively.

   (a) *Spatial segmentation*: Spatial segmentation on videos is similar to image segmentation applied to individual frames.

   (b) *Temporal segmentation*: Temporal segmentation partitions the video into shots based on visual similarity between consecutive frames [90], as shown in Fig 2.3. Some works consider scene-based segmentation, where a scene is semantically a higher-level video unit compared to a shot [90, 189]. Detailed information can be found in section 2.1.1.

   (c) *Spatio-temporal segmentation*: Spatio-temporal segmentation results in object trajectories, or temporally linked spatial segments over multiple frames. The latter can be achieved by color or motion tracking or inter-frame segmentation [313].

Feature extraction:

Feature extraction aims at encoding the content of the multimedia items in a concise and descriptive manner for use in retrieval, recommendation or similar systems [221]. Feature extraction typically is a multi-disciplinary research field and depending on the application at hand, it can requires knowledge from various fields. For instance, since ACA deals with audio signals, the emphasis lies on digital signal processing, and depending on the tasks, it can also require knowledge on machine learning, musicology, music theory, even (music) psychology, psychoacoustics, and audio engineering [199]. An accurate feature representation can reflect item characteristics from various perspectives and can be highly indicative of user preferences thereby crucial for RS.

Features can be represented by manifold data types, most commonly as a scalar describing a single aspect (*e.g.,* beats per minute of a song), as a high-dimensional numeric feature vector (*e.g.,* values of a color histogram for the three channels red, green, and blue), or as a set or list of feature vectors, where each individual vector describes one segment of the item under consideration (*e.g.,* energy levels in different frequency bands of an audio frame).

We identify four main categories of feature extraction techniques for recommendation purposes:

(a) *Type I: Unsupervised feature extraction:* This types of feature extraction refers to the ones extracting a feature vector from a multimedia item without the use of any previously human-assigned semantic labels during or after the extraction process or without taking into account the relation with other items in the dataset. The term "unsupervised" refers to such unlabeled multimedia documents.

The majority of feature extraction techniques fall into this classification; the main advantage of these techniques is their simplicity since a feature extractor can be taken item off-the-shelf and used on a MM item; the main drawback of such methods is that each feature vector represents the characteristic of a single target item without taking into account the inter-relation between items in the dataset. Examples of features extracted using this type of feature extraction scheme include MFCC and BLF for the audio domain or color and texture in the image media type.

(b) *Type II: Unsupervised feature extraction using an external multimedia corpus:* In this type of feature extraction, instead of using a feature extractor off-the-shelf and applying it to a multimedia item, an external multimedia corpus is used. The information obtained from the multimedia corpus is integrated for the final feature description.

The multimedia corpus can be also built by considering the items in the train dataset. For instance, in an offline settings, the extracted feature vectors from the training dataset can be used to build a statistical model of the entire multimedia collection. This statistical model is sometimes referred to as *universal feature model* or *audio/visual dictionary* whose role is to map the extracted features in a new feature space in which the interrelation between items is better modeled. This approach is opposed with the previous one which disregards the relations between items and treats them as individual entities. During test time, both the model used in the train phase and the feature extracted are jointly considered in order to build the final feature vector. Examples of this type of feature include i-vectors [82] in the audio and Fisher vectors in the image domain [10, 248].

(c) *Type III: Supervised semantic-inferred feature extraction:* This class of feature extraction techniques uses some form of human-generated semantic labels assigned to multimedia items in the post-processing stage of features extracted and from this perspective it can be considered as a supervised approach, since human knowledge is embedded in the final feature descriptions. Examples of such semantic-inference is learning the association between low-level multimedia content (e.g., music) and emotion, genre and or movie popularity. Linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (pLDA) are techniques that are commonly applied to transform the features into a new semantic space where the ratio of between-individual variation (intra-class variance) to within-individual variation (inter-class variance) is maximized. The models seeks directions in space that have maximum discriminability under a target label *e.g.,* genre separation or emotion classification. Such methods are very suitable for class recognition tasks [253].

In an offline setting, a model which contains the association between low-level multimedia content and the labels is learned denoted and exploited during test phase to map the extracted features into a new semantic space.

(d) *Type IV: End-to-end (E2E) learning:* End-to-end learning often refers to training a possibly complex learning system by applying gradient-based learning to the system as a whole. This technique has been popularized in the context of deep learning [129]. While the previous class of techniques treats the feature extraction and recommendation as separate stages, in E2E learning, all intermediate processing steps are optimized simultaneously. For a recommendation task, E2E learning encompasses employing deep learning methods that accept as input the segmented multimedia file (*e.g.,* video frames) and the parameters at all layers are jointly trained to optimize (*e.g.,* minimize) an objective function for the desired task.

Feature post-processing:

This stage involves applying different processing steps to the extracted features mostly with the objective of creating an item-level descriptor by aggregating the features over time (temporally), early fusion, or dimensionality reduction. As for temporal aggregation, the following approaches are widely used in the field of multimedia processing:

(a) *Statistical summarization:* The simplest approach to build an item-level descriptor from the computed features is to compute a statistical summary of the local features that make up an item, using standard statical functions such as mean, standard deviation, median, maximum, or combinations thereof, *e.g.,* means plus covariance matrix. A subnational amount information may be ignored using this approach. In addition, the temporal ordering of the features is discarded when using statistical summarization [180].

(b) *Probabilistic modeling:* An alternative approach for temporal aggregation is to summarize the local features of the item using a probabilistic model. Gaussian mixture models (GMMs) are often used for this purpose. A GMM describes the item under consideration by a fixed number of Gaussians, given by their parameters (means and covariances) and mixture weights. Formally, a GMM is characterized by a number K of Gaussian distributions $\mathcal{N}_k = 1...K$, each of which is defined by their means μ_k and covariances Σ_k. Each Gaussian $\mathcal{N}_k$ in the mixture model is further assigned an importance (mixture) weight w_k, which approximates the number of local features represented by $\mathcal{N}_k$. The parameters of a GMM are typically learned by maximum likelihood estimation method (MLE) [180].

(c) *Other approaches:* Other feature aggregation techniques include vector quantization (VQ), vectors of locally aggregated descriptors (VLAD) and Fisher vectors (FV), where the latter was originally used for aggregating image key-point descriptors. They are used as a postprocessing step for video representation, for example within a convolutional neural network [128].

The fusion of multimedia features is an open and promising research direction to improve performance of multimedia tasks, such as recognition, classification, or retrieval. The fusion at the stage of feature postprocessing is known by *early fusion* whose goal is to combine feature extracted from various unimodal streams into a single representation (see [304] for a discussion on early v.s. late fusion). For instance, in [289], six different *block-*

level features (BLFs) capturing spectral aspects, rhythmic aspects and tonal aspects are proposed and fused together using an early fusion method. The appropriate synchronization of the different modalities, specifically when and how much data should be processed from different modalities, is still an open research problem. A thorough discussion about this topic can be found in [31].

Content-based learning of user profile:

The goal of this step is to learn a user-specific model based on her past history of interaction with the items and descriptive attributes of items (item feature vectors) which is leveraged in the next step to predict the target user's preference on the items [19].

Filtering and recommendation:

In this final step, the learned user profile model is compared to representative item features (or item profiles) in order to make recommendations personalized to the target user.

2.2 State-of-the-Art Approaches

In Section 2.1, we presented an overview of the main concepts in multimedia processing and recommender systems as well as our framework to characterize a complete content-based multimedia recommendation system, which has the primary goal of providing personalized multimedia recommendations to users by exploiting multimedia content. In this section, we present a systematic literature review of recommender systems which exploit multimedia content for a particular media content recommendation (music, video, image and text) but also in domains where the target product is not multimedia necessarily (*e.g.,* place of interest (POI) in tourism). This literature review discusses the state of the art and latest advances in the field in the following directions:

(1) We present a classification of the *main research challenges and goals* addressed in the literature for the research works that have exploited multimedia content information in music, image, or video recommendation.

(2) Based on this categorization, we provide a description of each research work and discuss its advantages and limitations.

Table 2.5: *Classification of main challenges addressed in the literature. Contextual variables include: mood/emotion and situation, location/POI, multimedia context)*

Goal/Challenge	Research work
Improve recommendation quality	[104, 293] [341–343]
Popularity bias and long tail	[63] [285]
Multimodal analysis	[234, 315] [196, 229, 339]
Context-aware multimedia recommendation	[73, 172, 201] [66, 72] [229, 266, 339] [139, 348]
Leveraging social network information	[53]
Sequence-aware multimedia recommendation	[162]
Unobtrusive Preference Identification	[336]

We categorize the reviewed state-of-the-art research works with respect to the objectives and challenges they address. The summary of main categories are outlined in Table 2.5 and respective research works are discussed in the following.

2.2.1 Improve recommendation quality:

The research works discussed in this section follow the broad objective of improving the recommendations quality *w.r.t.* particular evaluation metric(s). These research works use various multimedia descriptors as well fusion schemes in order to achieve this goal.

As one of the first works in the music (audio) domain, [104] proposed a music recommendation system which combines item-based CF and acoustic CB in order to build a hybrid recommender systems using feature-level hybridization. The proposed system in [104] takes a user's playlist as the input to the system and provides a list of recommended songs which are similar to those in the playlist. For a tutorial on this subject, find the Recommender System challenge 2018 [277, 278] as part of the challenge series at the ACM conference of recommender systems [11, 12].

The CF recommender works by building a co-occurrence matrix from song co-occurrences in playlists and decomposing the matrix using eigenvalue decomposition. On the CB system, three kinds of audio features are extracted: timbral (MFCCs), rhythmic, and time signature features (autocorrelation)[6] using the software package in [330]. Multiple acoustic features are concatenated to make an acoustic feature vector of dimension 30 (19 timbre, 6 rhythmic and 5 pitch). Afterwards, the CF spectral and CB acoustic features are combined by a linear weighing scheme which accounts for features that have significant correlation in a certain feature direction when calculating similarities between two music pieces. The advantage of the proposed hybrid system is that it leverages both social/cultural aspects of music together with it acoustic properties and is able to recommend more popular songs if it is fed with playlists containing more popular tracks or acoustically similar musics if fed with low co-occurrence songs [171]. The main limitation of this [104] research work is that it lacks an experimental study.

[342, 343] proposed a recommendation strategy whose origin can be traced back in [341]. The proposed system aims to solve the traditional trade-off between accuracy and diversity of recommended songs. The authors used *variety of artists* of the recommended music pieces as the main measure to compute the diversity of recommendation lists. The main issue reported by the authors is that recommendation generated by CF models are less diverse since most users provide higher number of ratings to songs associated with their favorite artists therefore limiting the diversity of recommendation lists. On the other hand, while CB methods suffer less from this matter, they often provide less accurate recommendations (compared with CF) since they solely rely on musical properties properties of songs in order to predict interesting pieces. The solution proposed by the authors is a hybrid recommender system that leverages the advantages of both CF and CB models via a feature-level hybridization scheme. The main novelty of the research works is on proposing a hybrid music recommendation algorithm based on probabilistic generative model named *three-way aspect model* which can explain the generative mechanism of the observed data by introducing a set of latent variables (called conceptual genres). The content-based audio features used in [341–343] are based on Mel-Frequency cepstral coefficients (MFCCs) and includes 13 coefficients of MFCC and GMM with number of mixtures equal to 10 and 64. Higher number of mixtures is reported not to make a significant difference in recommendation quality. The CF approach is based on a rating matrix using a 3-point rating scale: dis-appreciation (0), neutral (1), and ap-

[6] Note that the CB system is infact a hybrid system by itself since it combines several content features capturing different musical aspects.

preciation (2) [60]. The main advantage of Yoshii et al.'s works [342, 343] is on proposing an *incremental learning* method that can handle increasing numbers of users, permitting new users to register to the system in an incremental fashion and at a low computational cost, without trading off accuracy. However, the system cannot incrementally add non-registered pieces (new items) to the probabilistic model [343]. However, these research work seems to be lacking an evaluation of whether semantic properties of latent variables are similar to those of existing genres or moods.

2.2.2 Popularity bias and long tail

An open research problem with RS is the issue of *popularity bias* (*aka.* long tail property or concentration bias). According to this property, the rating distribution of users among items is highly skewed in which a much larger portion of items reside in the long tail (very low rating frequencies), while a much smaller portion reside in the short head (very high rating frequencies). These two categories of items are quite often referred to as *niche products* and *mainstream (popular) products*, respectively. From a business perspective, most business profits arise from products that stay on the long tail while those in the short head entail little profit. On the other hand, from a user perspective, recommendations of items that the user did not know about is an important characteristic of RS, because it can lead to discovery of new items which is one of the main pragmatic goals of such systems [23].

The popularity bias problem refers to the fact that majority of the recommendation algorithms have the tendency to promote recommendation of popular items and therefore recommendations generated by such system suffer from lack of novelty [25, 78]. For instance, it is known that item-based and user-based CF base their recommendations on frequently co-rated items. Hence, recommendation of less rated items by such algorithms can be misleading [25]. Novelty in RS is interpreted as the likelihood of a RS to generate recommendation that a user was not aware of or did not know about. A good discussion of the notion of novelty and ways to measure it is provided in [187, 292]. Since users usually provide solely some form of preference (either implicit or explicit) to items, it is difficult to understand novelty by only considering these preferences. Therefore, the most natural way to understand novelty is through user-studies and directly asking users whether they were aware of an item previously or not. In offline studies, the standard approach to estimate novelty is by computing the average self-information of recommended items, which is equivalent to say novelty of a recommendation list is proportional to how much "unpopular" the items are [317, 353].

In the following, we review a number of research works that use some form of low-level multimedia content to address either of the following related problems:

(i) popularity bias as a standalone problem and its roots

(ii) novelty *vs.* accuracy of recommendation

In order to address the above research questions, typically a combination of CF and CB recommenders based on multimedia content is used. This is because these two family of algorithms base their recommendations on inputs of entirely different nature. For example, CF models traditionally show to suffer from popularity bias because of their dependency on user-rating to compute recommendation. On the other hand, multimedia CB models which base the recommendations on the multimedia content, treat different items more equally and less promote recommendation of popular items.

In [63] the authors present an interesting study with the main objective of investigating if the network topology of item-item CF has any pathology that hinders recommendation of novel items. This study stands at shark difference with other studies that solely put their concentration on comparing the ability of different recommendation models in computing novel recommendations. For this reason, the authors present an offline study in which the primary goal is to analyze the topological properties of the item graphs connecting recommended items. The authors conclude that social CF recommenders in Last.fm are susceptible to popularity bias. The main limitation of the offline study is that it does not reflect information about the items users have consumed prior to using the RS, thereby limiting the computation of novelty [25, 292]. To remedy this issue, in a second study, [63] presents a user-centric experiment to assess the perceived novelty of recommendation from the users' perspectives. The recommender takes the 20 most played artists in the target user's playlist as the input and uses an item-item CF together with acoustic CB to generate recommendation. The CB system uses MFCCs audio features and features related to rhythm and tonality (key and mode). The authors use *artist similarity* as the measure of novelty. The results obtained by the user study with 228 users and 5,573 tracks indicates that even though CF recommends less novel items than CB, the perceived quality of recommendations is higher in a social CF compared with an acoustic CB. The main advantage of this [25, 292] research work is that it conducts two types of studies (offline *v.s.* user-study) in order to draw a conclusion about different recommendation models. Although user-studies are difficult to conduct in reality, the approach adopted by the authors is promising and adds significant value to the final conclusions. One of the shortcomings of this [63] research work is that it does not well elaborate how to use these results with the main goal of the recommender system, *i.e.,* optimizing revenue and recommending useful items [132].

[285] followed the study by [63] from a different perspective. The authors mention that generally there exists two main reasons why niche products stay hidden in the long tail: (1) lack of reliable information about a niche product so the niche products are recommended in a wrong context where users most likely are not interested in it, (2) a niche product is hardly or not at all reachable by any sequence of recommendations but only by a direct query. While most previous research works including [63] focus on the former problem, the authors of [285] focus on the latter problem. The main question the authors studied was whether recommendation networks generally suffer from reachability problems. They tested two recommenders, an item-item CF and an acoustic CB, where MFCC features were used in the latter. Experiments with two recommender types *w.r.t.* coverage show that in music recommendation networks, a fraction of items remain unreachable independent of the recommendation approach and the network size. This is an interesting result and is known as the problem of *concentration bias* of recommendations. An equivalent concept in other research areas is *the hubness* [256], where "hubs" refer to points appearing in a high number of k-NN lists, effectively making them "popular" nearest neighbors [256], without actually being similar.

[285] can be considered as one of the first research works studying this problem in music recommendation, however its main drawback is that it does not provide a practical solution for the hubness problem. Follow-up works such as [181, 284] propose approaches to alleviate hubness by rectifying the item similarity space. The authors show that hubness reduction in fact increases nearest neighbor classification accuracy on a variety of machine learning datasets as well as for the task related content-based music recommendation. Techniques like data normalization has been shown to improve retrieval accuracy independently of the underlying method.

The proposed system in [293] seeks a somewhat similar question to the previous research works that is to find an effective strategy to solve the trade-off between accuracy and *novelty* of recommended musical pieces. Similar to [343], the authors use *variety of artists* of the recommended piece to quantify novelty. The main challenge reported in the work is that music RS typically use two popular functions for measuring music similarities, the (unweighted) cosine similarity and the (weighted) Minkowski distance.

According to the authors, these similarity metrics are not sufficiently proper to uncover the differences between "evolving" and "dynamic" nature of musics from (1) one type of music piece to another type or (2) a user with particular music preferences to another user. For example, the Minkowski distance assigns a static weighting scheme when determining the similarity of music. This can be an inappropriate choice when feature weights vary from one type of music to another. For instance, as mentioned by the authors in [293] in some music genres such as rock music, the audio intensity can be an important feature, but it can be much less important in determining music similarity for classic music, making it necessary to design a dynamic (time-varying) weighting scheme to different acoustic features when moving from one type of music piece to another. In addition, the perception of the same pieces of music is different for different users [276]. Therefore, it is also necessary to use dynamic weights to every audio feature in order to capture the subjective differences between users. The proposed solution by [293] is a *metric learning* approach which learns appropriate similarity metrics based on the correlation between acoustic features and user access patterns of music. The system leverages hyper-graphs in order to combine usage data and content similarity information. As such the main advantage of the proposed approach is that it uses a variety of information about users, including access patterns and listening behavior, together with acoustic properties of music pieces, for user modeling and dynamic optimization. A major limitation of this method is when two music pieces have significantly different feature values, but are appreciated by the same group of users, the accuracy of the weights computed may be less precise [209,306].

2.2.3 Multimodal analysis

A multimedia document can be viewed as a result of an *authoring* process as illustrated in Figure 2.5. The author *e.g.,* the film director, is the creator of the media content and has sufficient knowledge about the domain. He applies various professional conventions to combine media contents (such as music and images) in order to produce a multimedia document in a process called *the multimedia authoring process* [302, 303]. As it can be noted, the author communicates with the user along different communication channels, the visual, auditory, and textual channels in order to convey messages, invoke emotion and so on [303]. Hence, the content of the multimedia item can be intrinsically multimodal. On the other hand, in order to make the multimedia document readable and understandable by machines, some characteristics of the media need to be extracted from a multimedia document and represented in numerical feature vectors. This process is called *multimedia content analysis* and is a critical step to mark the transition from the qualitative to a quantitative presentation. The interpretation of the media by the media consumer depends on several user-related factors, such as her *demographics* (age, gender, *etc.,*), her *emotional state*, *personality*, and *context* (time, location, the device she is using to view the multimedia content) to name a few. The user expresses her opinion about the content by an implicit (purchasing) or explicit (rating) action.

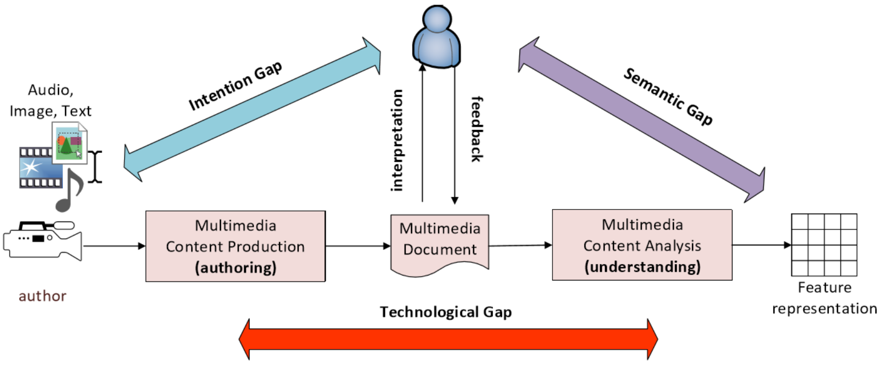


Figure 2.5: *Multimedia authoring process. The figure is a modified version of the one proposed in [302], tailored for a recommender system using multimedia content*

A multimedia RS using multimodal analysis can provide effective personalized recommendation when:

(a) The perception of the media by the user corresponds well with the intention of the author, in other words there exists no or little *intention gap*. This aspect is out of the RS designer's control and is more related to the expertise of the author and user-specific factors (experience, demographics *etc.,*) driving her perception.

(b) The feature representation of the multimedia document by the machine matches well with the interpretation of the media consumer from the media, *i.e.,* there exists no or little *semantic gap*. This aspect is a known and highly addressed problem in the field of multimedia information retrieval [166,217] particularly and is associated with the gap between the semantically high-level queries that the user provides as the input and the low-level feature representation used by machines used to retrieve media content. The notion of query in information retrieval is similar to the user profile in RS; however it remains a question from a semantic expressiveness point-of-view if the user profile should be considered low-level or semantic or combination of both.

(c) The feature representation of the multimedia document by machines is as much similar to the stylistic and semantic conventions by the author to create the media, *i.e.,* there exists little or no *technological gap*. This latter aspect is a machine-related notion and relates to the extent the machines today are intelligent in representing and reflecting the conventions made by the authors.

In multimedia RS, we can use different modalities to build a multimedia content-based RS. For instance, in music RS in addition to the audio signal, metadata in variety of form can be used namely: reviews (textual), lyrics (textual embedded in audio signal), album covers (visual), music videos (visual) and surrounding text to a music file in Web page. All these information sources serve as content description for a music piece and can be leveraged for recommendation.

[315] proposed a hybrid CB + item-based CF by merging the two systems using a simple ensemble-based hybridization scheme based on template-based combination rule. The CB system exploits metadata (genre, year and mood) and acoustic features (timbre and tempo) and merges them at similarity level using a linear combination of similarities with equal weights: $\mathrm{s} = \mathrm{s}_{genre} + \mathrm{s}_{year} + \mathrm{s}_{timbre} + \mathrm{s}_{tempo} + \mathrm{s}_{mood}$ [321]. Preliminary results of recommendation quality with regards to MAE indicate that the ensemble hybridization improves the quality of recommendation compared to individual CB and CF recommenders when used in isolation. Ensemble-learning methods have been widely adopted for building hybrid RS and have been used in grand challenges such as the Netflix Prize contest and KDD Cups [312]. The main limitation of this research work is that very little information about the fusion process is presented and evaluations are carried out *w.r.t.* a simple error metric only (MAE). In addition, the authors make no attempt in merging the similarities in a more effective manners (for example using a linearly weighted approach and possibly learning for those weights).

[234] proposed a recommendation approach that connects users with items by considering both ratings and social tags. It is mentioned that tags are rich multifaceted source of information and contain information about genre, style, instrumentation as well users opinions and emotions. Therefore, incorporating tags into a unified recommender model can provide more accurate and personalized recommendation. The proposed system in [234] is called MusicBox which automatically assigns tag to music

by capturing a three-way correlation between users-tags-music items using three-order tensors. CB audio features are leveraged for tagging of the items that are not tagged previously by users. The tensor factorization technique, just like matrix factorization, decomposes the input tensor into multiple factor matrices and a core tensor [298]. The main advantage of this [234] research work is that it integrates rich tag information in a novel recommendation strategy, to improve recommendation accuracy. However, other extra useful information (such as user context) can be also considered into this model to improve its effectiveness.

[229, 339] proposed an interesting systematic framework for video recommendation by leveraging multiple of modalities (audio, visual and textual). The proposed system is called VideoReach in which it characterizes a video by a set of features extracted from all the constituting modalities, such as audio, visual, and textual (text inside the content), together with metadata. The textual and metadata considered in this research work include the query, keywords, and surrounding text, as well as (automatically generated) transcripts. The similarity between two items is described as multimodal relevance, that is the combination of textual, visual, and aural relevance. The authors propose a mechanism to leverage the relevance feedback from users to adjust the inner-modality (the weights associated to features in one modality) and inter-modality (the weight associated to each modality as a whole) fusion weights.

As for the pros of these [229, 339] research works, the proposed system integrates multimodal relevance and user feedback in a contextual video recommendation which does not require a large collection of user profiles as the current video watched can be used as the input to the system defining a new form of context called *multimedia context* [171, 216] (see next section). This is while most conventional RS heavily rely on a sufficient collection of user profiles (associated with different users) to effect recommendation. Other similar recommendation algorithms which recommend video *w.r.t.* a seed video that the user is currently watching include those based on graph representations, where the seed video is the root node and connect to every related video through an edge. An example of such graph-based representation is the *Related Video Graph* (RVG), which is described for the YouTube recommendation system [35, 352]. Another advantage of [229, 339] is that it is one of the pioneer works in using multiple modalities (audio, visual, textual) in a building CB movie RS while majority of previous works rely on single modality (*e.g.,* visual or metadata).

As for the limitations, the three modalities in [229, 339] are combined using *predefined* relevance scores, combined in a linear fashion. Some previous knowledge has been assumed giving textual relevance a higher importance compared to visual and audio modalities. No learning is involved for computing (defining) these weights. In addition, the multimedia features extracted are quite old and deprecated. Finally, as mentioned by [271] using video features as described in [229, 339] might not be worthwhile since the calibration effort can be too expensive.

2.2.4 Context-aware multimedia recommendation

The notion of context has attracted significant attention in RS research, specially due to the advancements of network and mobile services and the growing tool and device landscape, with a wide range of sensors are packed into todays "smart devices". As a result, a new paradigm of recommender systems has been developed in recent years to demonstrate the potentials of contextual recommendation. Such system are commonly referred to as context-aware recommender systems (CARS) [320]. However, the term "context" is defined rather vaguely, and its definition also differs between different communities, even between different researchers in the same community. In [37], the authors provide a good review and categorization of 36 context models and context types used in research on intelligent systems. One of the earlier definitions, which is broad enough to still hold today, is given in [102]: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves."

In the domain of RS, context typically describes the situation of the user when searching for items (*e.g.,* time, mood, and current activity). Clearly, such information can influence the information or entertainment need of the user and thus should be taken into account in addition to the information about the user's long term preferences, when providing recommendations.

The contextual information can be categorized in several ways. As one of the old works [282] classified context based on: *computing environment* (available processors, network capacity), *user environment* (location, social situation), and *physical environment* (lighting and noise level). [13] categories context into *primary level* which can measured directly (location, identity, activity, and time) and *secondary level* (such as emotional state of the user). Charu [20] classifies context into a simpler categorization based on *time*, *location* and *social information* (user's friends, tags, and social circles). Finally [171] provided a well-categorized classification of contextual factors as: *item-related* (location, time, weather, other parameters such as traffic, noise level or traffic jam), *item-related* (activity, demographic and emotional state) and *multimedia context* (text, image, music, video). We will use this classification scheme in the description of the following content-aware systems.

Mood, Emotion (user-related)

Moods and emotions clearly have a direct impact on our thoughts and preferences. Although a clear definition of emotion has been proven elusive [161,232], emotions have been described as internal mental states characterizing reactions to events, agents or objects that vary in intensity [232, 240]. Emotion and moods are different from each other in the sense that emotions are considered as short-lived, intense response to some external stimuli whereas moods are untargeted, longer enduring experiences. According to an old theory, namely mood management theory (MMT) [355], people use media to modulate their affective states and try to alter negative moods and maintain or prolong the positive ones. In other words, due to hedonistic desires, individuals organize their surroundings to adjust a full scope of moods using specific genre or specific of form of communication available [232]. While many scholar agree with MMT, the theory has some limitations as well since it cannot explain some paradoxes of media selection. For instance, the pleasure of watching horror movies seems to be in contraction to MMT's assumption of hedonism [238]. Another theory named mood adjustment believes in mood optimization that is more general than mood regulation and is articulated by the fact that people consume media to achieve the mood that they believe to be most useful [183]. Finally, a recent theory states that people's desire to consume media are not driven by hedonism rather eudaimonia which is happiness rooted in greater insights and connections to human experience. This theory can better explain individual's desire to use tragic or horror movies [239]. Further information can be found in [232]

Regardless of which theory is the most accurate, almost all related-theories to date agree on the relation between emotion and media consumption. From a RS designer point-of-view, this implies that when someone is depressed for a particular reason, she may want to listen to a piece of music that can cheer her up as an example. At this moment, she will search music by mood most likely with disregard to what the melody sounds or whom the artist is [99, 167]. Traditional MIR systems do not allow meeting users' various needs and to search music via their desired mood and as such there is a strong requirement for building new multimedia recommender system that uses emotion in understanding multimedia [99, 171]. In contrast to

CF approach which recommends multimedia product based on similarity with other users, content-based approaches which base their prediction on perceptual attribute of multimedia (*e.g.,* pitch, intensity for music or color variation, motion for video) can characterize the expression of multimedia emotion (*e.g.,* music or movie emotion). For example, rapid tempos with relatively constant pitch is often characteristic of a happy music. Multimedia content analysis together with advances of machine learning allows us to predict the dynamic perceptual attributes of multimedia directly from the raw signal and use them in design of personalized affective recommendation systems [99]. In the following we review research works that apply emotion and features extracted from multimedia content for the recommendation services.

[196] proposed a generic generic emotion-based music recommendation model to recommend music by discovering the relationships between music features and emotions from film music. Potential applications of emotion-based music RS as stated by the authors include but not limited to recommending music that matches well with a home-production video, setting up background music of shopping mall, context-aware home for accommodating inhabitant's emotion, and music therapy. The authors use acoustic features extracted from film-music including: *melody*, *mode*, *tempo* and *rhythm* together with an affinity graph to discover the association between music features and emotions from film music. Given the query emotions, the recommendation model will return the recommended music features with respect to the query. The recommended features are then employed to rank the music dataset and to recommend music for a query emotion. The evaluation of the approach was carried out with a collection of 107 film music from 20 animated films. The results showed that the proposed approach using the recommended features can identify the emotions of a music track with an average accuracy of 85%, however no comparison with existing emotion recognition approaches has been performed.

This research working carries some interesting ideas for advancing the research on emotion-aware RS, however it has some limitations as well. For example, the authors did not use a particular emotion model but relied on a set of emotion presented in [260] with some additional manually inserted emotions such as lonely and nervous to create 15 emotion labels for their tasks. In addition, as pointed out in other research works [241] emotions should not represent only media characterization and a complete system should also consider user-generated tags about the content of the media.

[65] proposed a contextual emotion-aware system to propose music pieces to a user based on the emotions predicted from the article she writes.

The rationale behind this research work is that people sometimes express their emotion toward a music by writing about them (article, free keywords). In the light of this observation, user-generated content can be seen as a rich source for finding emotional context information. The proposed approach entails modeling the relationship between user-generated text and the music listening behavior by using factorization machine (FM). To this end, the system uses the listening history of the user with the CB audio information extracted from her music tracks and combines them with the contextual emotion information mined from user-generated articles to improve the quality of music recommendation. The audio features are collected from the EchoNest website [7] and include loudness, mode, and tempo, danceability of music while the textual features include term frequency and inverse document frequency. The work also uses ANEW affective lexicon [48] in order to generate affective features that characterize the user's emotion state from text.

A less investigated area of research is emotion-aware recommendation models for other multimedia domains such as images and videos. [201] proposed a system to automatically suggest music based on a set of images (paintings). As the motivation, it is stated the affective of content of painting when harmonized with music can be effective for creating a fine art sideshow referred to as *emotion-based impressionism sideshow*. Emotion is used as the main "catalyst" to enable the association between the painting (the input) and the music (the output). The proposed method extracts a set of visual features such as color, light and textures from the paining images and discovers the emotion of the paintings by training using the Mixed Media Graph (MMG) model, originally used for finding correlations between multimedia objects [244]. In a similar manner, the correlation between music track features and emotions are discovered by training the MMG on a manually labeled training set of tracks. The acoustic features used are melody, rhythm, tempo from the music since these musical elements are known to be affecting emotion. The method employs the ROCK algorithm [131] to cluster paintings into groups and then a music clip is recommended based on each cluster's common emotions. The system further proposes a composition step whose goal is to "schedule" the display so that the dissimilarity of affective content between the adjacent slides is as low as possible. The Earth Mover Distance (EMD) is used to measure the similarity between painting since correlated with human perception of texture. The editing procedure is then followed by some additional steps such as close-up animation, synchronization and rendering. For the evalu-

[7] (http://echonest.com/)

ation, the authors used a dataset consisting of 138 impressionism paintings and 160 classical piano compositions. Evaluation is done subjectively by asking 18 users to view different slideshows and express their audiovisual satisfaction. Music recommendation based on a set of images, or generally identifying the association between two heterogeneous modalities such as images and music, has been largely under-explored and this can be considered as a merit for this [201] research work. However, this work has some serious limitations as well some including: the authors mention that the labels were Russell's circumplex emotion model, however the exact emotion labels used to mark music and paintings remains unclear. In addition, the work does not present a clear description of the evaluation procedure and their comparison baseline [172]. Furthermore, the proposed approach does not address a general set of images (*e.g.,* photographs) and music track which can be a potentially relevant and interesting task as addressed in [308].

[266] proposed a mood-aware music recommendation system that computes music similarities by their emotions. Specifically, first emotional information from music acoustic properties (scale, rhythm, harmonics) are extracted by training a SVR-based mood classifier. As the result, each music is represented as emotion vectors in Thayer's model from which music similarities are computed. For recommendation, two types of models are used (1) an item-item CF and (2) an ontology-based where the ontology-based recommender is used to infer user's mood from context information (time, location *etc.*). By comparing the inferred mood with the predicted mood of the song from the music content, recommendation is effected.

One advantage of this research work is that it improves the categorical labeling of emotions such as happy or say in earliest works on mood-aware music RS [117] by using a dimensional model of affect proposed by Russell [272] in which each emotion is described as points in a 2D plane spanned by the arousal and valence axes. The main disadvantage is that it treats the emotion as a static concept. This is while in most music pieces the emotion is dynamic and time-varying. In fact, composers often contrast different emotions, *e.g.,* lively versus calm for different music genres (*e.g.,* classical). This calls for an emotional model which is continuous both in time and in value [329].

Addressing some of these limitations in [139] the same authors propose a content-based music recommendation system with a distinctive difference from [266] by proposing an emotion state transition model (ESTM) to model human emotional states and their transitions by music. ESTM acts like a bridge between user emotion and low-level music features. For example, consider the case when a user feels "sleepy" while looking for an "exciting" music to shake her music. In recommending the right music to this user, the system should take into account her situation since the same exciting music which can be listened in a pub or at home cannot be recommended to her in a library or at night where soft or easy listening music are better solution to change moods.

The authors of [139] mention a number of theories in the literature to represent human emotion such as Russell [272] and Thayer [314] which are the two models representing a target emotion on a 2D space based on the limited number of cognitive components such as arousal and valence. Component process theory is another theory which characterizes human emotion being composed of fundamental cognitive components [281]. Motivated by this theory the authors expressed human emotion by the combination of emotion adjectives and their strength and state that at the same time each emotion can contribute to the transition to next emotion. The authors further propose a COMUS ontology for evaluating the user's desired emotion state based upon the user's situation and preferences. The music recommendation is effected based on the *current* and *desired* emotions and is accomplished with a music classification algorithm based on low level features.

A limitation of this work is that the user is required to select a situation from a *limited* predefined situation list. This together with other features such as age, occupation, hobby define user's overall contextual information. It is not possible to insert a new situation in the list [158]. Also, the ontologies do not consider the weight of emotional tags used [177]. The other limitation of this research work is that it assumes a prior explicit knowledge about how a specific individuals changes their music habits depending on listening context [328].

Location, POI (item-related)

In the recent years, location-aware music RS have been gaining attention due to their wide range of applications. With the technique, user's favorite songs can be automatically identified and retrieved based on where the user is present.

[172] proposed a CARS that suggests music for a given place of interest (POI). The systems accepts as the input location of the user and recommends her music that corresponds with her location. For instance, during a visit to the city of Salzburg[8] the user may appreciate a piece of music

[8] the birthplace of famous composer Wolfgang Amadeus Mozart and home for the musical play and film The Sound of Music in mid 20th

by Mozart sonata. Matching of music and POI in [172] is realized by: (1) representing items in two domain (music and POI) with a set of common tag features assigned by users reflecting their emotion toward items in two domains. The choice of emotional tags is motivated by the fact that both music and user POI can invoke emotions and the commonalities can be leveraged to establish a degree of matching between them [49]; (2) establishing a knowledge-based system by extracting information about two domain using semantic Web and its implementation in DBpedia Linked Data repository. The final music recommendation system is a hybrid combination of a "knowledge-based" and "auto-tag-based" recommenders, in which two systems produce a music-to-POI score. A rank aggregation method based on Borda count is used to combine two list of ranking produced by two recommenders. The auto-tagger uses musical properties of the signal and recommends a set of tags (with a fixed maximum size) with a given probability. It uses two set of audio features features: MFCC modeling the timbre and block-level spectral features. Two type of features are concatenated together to form a super-vector defining each track. The evaluation is performed in an online user-study where the users are invited to access the appropriateness of the music recommended to 25 POI. The results of the user-study shows that personalization of music via musical properties is not sufficient and it is important to implement effective adaptation techniques to the user's context. As one of the merits for this research work we can name successful usage of,the Borda count aggregation method is shown to be effective in similar tasks, such as group recommendation [34] and seem to be an appropriate hybridization technique for combining recommendations. A limitation of this work is related to the dataset since majority of the prior methods have only been evaluated based upon synthetic, small-scale and controlled-environment datasets; no effort has been taken for instance to address these problems using a real-world dataset that records people's music-listening behavior in a smart phone context and this can remain in contradiction with the research question that the personal, situational, and musical factors of musical preference [100]. Another interesting perspective is that the proposed method may not suitable for other domains such as location-aware TV show recommendation since the new TV show generally do not have any semantic tags and the video auto-tagging is not reliable [325].

Improving these limitations, [72,73] proposed a venue-aware music recommender system in which venue is defined as the place in which an activity or event takes place such as a library, gym, office or mall. The key difference between the research work [172] which we presented above and

[72, 73] is in the manner the linking between the location/venue and music is performed. While both research works exploit music acoustic content to represent music, the space in which the association between music and location is done is entirely different; in particular [172] uses "tags" assigned to music and user's location to find the association while [72, 73] map the characteristics of venues and music into a "latent semantic space", where suitability of music for a venue can be directly measured. The proposed system by [72, 73] is named *VenueMusic* and consists of two main functionality modules: a *Music Concept Sequence Generator* (MCSG) and a *Location-aware Topic model* (LTM) where the role of MSCG is to map the raw music stream into a sequence of semantic concepts while the LTM is exploited to represent songs and venue types in a shared latent space. The evaluation is done by using two datasets. The result of experimental study shows that (1) the MSCG has an overall acceptable performance in music-concept classification. The authors mention this as an important achievement since the MSCG plays a fundamental role in determining the effectiveness of the final recommendation model; this result supports the first research question on the fact that it is better to use latent topics to capture the associations between songs and venues compared to directly using low-level audio features or semantic concepts; (2) Is it better to represent songs as music concept sequences in the LTM than to represent songs as "bag-of-audio-words".

Multimedia context

As mentioned in the challenge section of multimodal analysis, it is common to combine a media content (e.g., music) with other media types (*e.g.,* text or image) in order to enhance information presentation, cross-selling of entertainment items, *etc.,* [171]. Therefore, the concept of multimodal context is closely related to the challenge multimodal analysis. The main difference is that the reviewed research work in this section, take as input a particular media content (image, music or video) and recommend as the output a set of other media contents or different form of information (POI).

Visual contextual advertisement is the very first related application field of multimedia context in which the content of particular multimedia item currently being consumed by the user becomes the target for recommending advertisements. The main goal here is to build a semantic match between two heterogeneous multimedia sources (*e.g.,* content of an image and the advertisement in textual form).

[348] proposed a visual contextual advertisement system that suggests the most relevant advertisement for a given image without building a tex-

tual relation between the two heterogeneous sources (*i.e.,* by disregarding the tags associated with the image). The authors mention that there exit two main approaches for visual contextual advertisement: (1) based on image annotation, (2) based on feature translation model. In the first case, a model is trained based on a selection of labeled images which is leveraged to predict text given a new image. Manual labeling of the items is required which makes the approach prone to error and labor-intensive. The second approach builds a bridge between the two visual and textual feature spaces through a translation model and then uses a language model to estimate the relevance of each advertisement *w.r.t* given a target image. Two main weaknesses are identified with SoA textual advertisement systems [70], first the term distributions of image tags and text advertisements do not correspond which result in absence of advertisement or irrelevant advertisements and second there exists a mismatch between text features and advertisements resulting in semantically irrelevant advertisement to the target images.

In [348] the authors propose a knowledge-driven cross-media semantic matching framework which leverages two large high-quality knowledge sources ImageNet (for image) and Wikipedia (for text). The image-advertisement match is established in the respective knowledge sources. The main limitation of the research work is that the proposed approach does not support semantic advertisement in an automatic fashion since new concepts are not added to the system automatically.

[66] proposed a system that generates several representative tags for each Flickr group by voting. The motivation is that many users would like their photos to gain public attention for social purposes and assigning photos to proper categories and attaching tags to them are proper approaches for receiving such attention addressed in this research work. The proposed approach by [66] consists of a concept detection method for photos and finding the best Flickr groups with the predicted concepts. The associated tags with these groups are thereby collected and ranked for recommendation. The main limitation of this research work is that ignores the role of user-preference (and personalization) in the tag recommendation process. Different users could have different tag favorites when exposed to a similar photo stimuli. Personalized tag recommendation can provide more appropriate recommendations results by taking the user profile into account [208]. Additionally, the method can only suggest tags from a predefined set. Finally, concerns have been expressed about the fact that the proposed tag-recommendation approach are purely content-based and do not leverage the collaborative knowledge shared between users [326]. The authors in [66] also proposed to exploit the predicted tags to search

for photo groups as recommendation. As the result, the proposed photo group selection in [66] heavily depends on the performance of tag prediction [207].

Another domain of interest for multmedia context-based recommendation is the *tourism domain*. Such systems can give the users a huge help in knowledge discovery and providing personalized recommendation to (mobile) users.

[71] proposed a personalized travel recommendation system by leveraging the rich and freely available community-contributed photos and considering demographical information such as gender, age and race in user profiles in order to provide effective personalized travel recommendation. In contrast to existing works which only use photo logs, the proposed system detects the above people attributes in the photos in an automatic fashion. It is shown that detection of such attributes are effective for travel recommendation - especially providing a promising aspect for personalization. For example, the authors discuss the correlation between travel patterns and people attributes by using information-theoretic measures. The main advantage of this research work is to focus on the importance of personalization in providing personalized touristic recommendation by leveraging rich visual data generated by users which are available freely. While a generic recommendation system, provides suggestion for a destination given by user by answering a question like: “I want to go to New York, what are the must-see attractions?” (CA mobile travel recommendation), the authors focus on two types of travel recommendation presented in the form of the following questions: “For a female, what is the suggested travel sequence in Milan?” (personalized route planning), “I am a female, I am now at Central Park in Manhattan, what is the next suggested destination?” (personalized CA mobile travel recommendation).

[226] proposed a recommendation framework for *fashion* that provides personalized clothes recommendation for a given clothes image by considering the visual appearances of the clothes. There exists two ways to offer clothes recommendation for a given cloth image: (1) finding some pairs of objects that can be seen as *alternative* to a query image (such as two pairs of jeans), (2) finding the ones which may be *complementary* (such as a pair of jeans and a matching shirt). While majority of previous works address the former task, the authors focus on the latter. Such systems are of considerable economic value and can be typically built by taking into account metadata, reviews, and previous purchasing patterns. The main novelty in [226] is the ability to examine the visual appearance of the items under investigation and to overcome some of existing systems’ limitations such

as the 'cold start' problem.

The proposed approach consists of extracting CNN visual features (pretrained on 1.2M ImageNet images) and performing and learning the visual similarity between a query image and the complementary items using metric learning. The main advantage of this work is to extend the typical form of clothes recommendation from the most similar scenario to the most complementary one which seem to be novel and useful by considering user needs and promoting business reviews. This work is also novel by putting attention on the visual appearance of items. As for the limitations, this research work is similar to a visual retrieval problem and differs from traditional recommendation system by disregarding users historical feedbacks on items as well as other factors beyond the visual dimension which can be important for offering useful personalized recommendations [149].

The recommendation model presented in [149] is called *visual bayesian personalized ranking* (VBPR) which uses a regression-based method similar to regression-based latent-factor model(RLFM) [18], defined by Equation 2.16 but also add a factorization term for uncovering the latent factors of users and items. The mathematical model of VBPR is given as

$$\widetilde{\mathbf{r}}_{ui} = \alpha + \mathbf{q}_u\mathbf{q}_i^t + \mathbf{b}\,\mathbf{f}_i^t + \mathbf{p}_u\,A\,\mathbf{f}_i^t \tag{2.18}$$

where α is the bias term which symbolically represents the global offset plus user/item bias terms, $\mathbf{q}_u$ and $\mathbf{q}_i$ are K-dimensional vectors capturing latent factors of user u and item i (respectively), $\mathbf{f}_i$ is the F-dimensional visual feature vector of item i and $\mathbf{b}$ is a visual bias term whose inner product with $\mathbf{f}_i$ models users' overall opinion toward the visual appearance of a given item. The role of term $A\,\mathbf{f}_i^t$ to transforms high-dimensional feature vectors (in this paper $F = 4096$) into a much lower-dimensional (say 20 or so). A is a $D \times F$ matrix embedding original visual feature space (F-dimensional) into the lower-dimension visual latent space (D-dimensional) and finally $\mathbf{p}_u$ is the user profile, a user-specific vector measuring the the extent to which the user u is attracted to each of D visual latent dimensions.

The authors apply the proposed VBPR on large-scale dataset from Amazon.com originally introduced by [226]. The datasets are about Women's and Men's Clothing but also consider Cell Phones&Accessories in which visual features have been shown to be meaningful. They additionally introduce a new dataset from Tradesy.com, a second-hand clothing trading community.

The main advantage of this [149] work is that it proposes a recommendation model to uncover the fashion-pattern among users and users' individual and global preference toward visual aspect of clothes in in a single model which can work effectively in cold-start scenarios. As for the cons, one may argue that although a single embedding matrix shared by all items can uncover the common features among different categories, *e.g.,* the characteristics which make people consider a t-shirt or a shoe to be 'colorful', however, different categories of products (clothes) may have subtle variations among them and a low- dimensional global embedding to capture these variations. To address this issue [147] extended this [149] research work by proposing a more flexible embedding using hierarchical a structures. In addition, some other research works [145, 148] have questioned the social and temporal aspect of the visual visual factors during user's decision making process and extended this [149] work along these dimensions.

2.2.5 Leveraging social network information

Online social networks contain rich content and context information which are valuable sources for mining different kinds of knowledge [206]. Due to the network or graph structure typically established between users in terms of "friendship" or "following" relations, they furthermore provide additional information that can be leveraged in RS research.

[53] addressed the music recommendation problem in music communities by exploiting social network information and acoustic properties of music. The authors mention that in typical music social communities, such as Last.fm, each user can make friends with other users, listen to their favorite music tracks, team up to make playlists, join specific groups, and use keywords to bookmark music tracks, albums, and artists. The resources (music tracks, albums, and artists) can have relations with each other, *e.g.,* a music track being part of an album. Typical audio content-based RS suffer from the "semantic gap" problem; CF models and CB ones using metadata have to deal with rating sparsity and noisy or insufficient data, respectively. The proposed approach by [53] is named music recommendation on hypergraph (MRH), whose goal is to learn a unified hypergraph to model multi-type objects and relations in social networks interested in music. A hypergraph is a generalization of an ordinary graph in which different relationships (*e.g.,* music, tag, and users) are modeled via *hyperedges*, which are capable of capturing high-order relationships. The authors empirically explore the contributions of different types of social network information for recommendation and show that MRH is helpful for practical music RS. In general, hypergraphs are widely investigated in information retrieval and pattern recognition tasks [125, 165, 210]. The main contribution of the approach proposed in [53] is that it constructs a hypergraph to leverage the

social media information by modeling the multi-type objects in a music social community as vertices and the relations among these objects as hyperedges. For example, acoustic CB similarities between music tracks can be treated as one kind of relation, friendship connections as another, and all these relations are combined in a unified framework. To sum up, the original aspect of this work is to spark new ideas on the development of novel models to capture heterogeneous data types and relations [210]. The model however seems to be complex for development in real application.

2.2.6 Sequence-aware multimedia recommendation:

Sequence-aware recommendation (SAR) is different from traditional rating prediction (aka matrix completion (MC)) in a number of ways [16, 255]. Mathematically speaking, let $g : U \times L \to R$ be the utility function which measures the usefulness (utility score) of a given sequence $l \in L$ for user $u \in U$. The SAR problem refers to determination of the sequence l which maximizes the utility score for the user u, *i.e.,*

$$\forall u \in U, \quad l_u = \arg\max_{l \in L} g(u, l) \tag{2.19}$$

In contrast to MC, the goal in SAR is not to predict the utility score for each user-item pair rather in computing an *ordered list* of items l with length k for each user, where each element of $l \in L$ corresponds to an element of $i \in I$. For a comparison of the SAR problem with general MC setup, compare the equations 2.8 and 2.19.

The main input to SAR problem is an ordered and often timestamped set of user previous actions (*e.g.,* anonymous users which are not logged in the system), where each action can belong to one of several predefined types (*e.g.,* viewing an item or making a purchase) and the users and items can have a number of additional attributes similar to standard MC setups. However, in contrast to MC, the preference score to items can be coming from unknown users in which case the relevant information is extracted from past anonymous sessions. In addition, the input data are not usually provided in the form of explicit ratings (as in MC) rather in the form of Web server logs.

The output of a SAR model is an ordered list of items in which not only the "ranking of items" is of importance, similar to a traditional recommendation setup, but also the transition from one item to another item (*i.e.,* ordering of items) within the recommendation list is highly desired. For example, consider the scenario where a user is recommended a rock music right after she has listened to a classic piece by understanding the fact that the user likes both musical genres. Even though both music styles are in agreement with her taste, the transition between songs and the coherence between song type plays an important role toward user satisfaction. In SAR, different recommendations are no longer considered as a set of different candidates for the users to select from rather the user is considered to be using the *entire* recommendations. The most common example of SAR is in music domain for music playlist recommendation while in other domains we can name video streaming as well recommendation of a series of video learning courses.

[162] proposed an automatic music playlist generation (MPG) method by recommending musics that not only appeal to the general taste of the listener, but also are coherent with the *most recently* played tracks. The authors mention the main particularity with MPG is that recommended items are consumed "immediately". Therefore, the SAR algorithm needs not only to take into the match between recommendation and the listener's overall preference but that the playlist is in some level homogeneous considering the artist, genre, tempo or smoothness of the transition. The importance of such multi-criteria recommendation is increased in applications like virtually endless playlists (radios), where the set of recommendations should not only be coherent within themselves but also with the most recently listened tracks.

Among different variation of MPG, the focous in [162] is on the immediate next-track recommendation (aka list continuation problem) given a history of recent tracks. The proposed method by this research work is two-phase method in which first, a set of candidate music tracks that seem more matching with the most recent listening history are selected by a multi-faceted scoring method that takes into consideration attributes such as track co-occurrences, musical and metadata characteristics and second phase, the first items of the list are re-ranked using an optimization approach that attempts to *minimize the difference* between the recently listened tracks and the continuation in terms of one or more desired quality dimensions.

An empirical evaluation in offline setting on three playlists dataset (Last.fm, AotM and 8tracks) shows promising results (in terms of accuracy by combining different information sources and the flexibility to balance between different quality optimization goals. The advantage of this research work is that it takes into account both short-term listening history (preference) and long-term preference of the user for recommending next song. In addition, a diverse and rich information sources such as the one based on CF, metadata features based on social tags and musical properties of the signal are take into account for generating the candidate songs. As for the limitations, the same authors in the follow-up work cast doubts [169] if computational measures used to evaluate the quality of next song recommendation match with the actual quality perception of music listeners and questions remains if the handcrafted playlist are actually reflecting real taste of many users.

Unobtrusive Preference Identification

Generating relevant recommendations to user requires knowledge about her preference. Both, the existing personalized search engines (PSE) and RS rely on user feedbacks of explicit or implicit to infer user intentions and preferences. Examples of implicit feedback for PSE and RS include the query history (PSE) or click data (both PSE and RS). In this line, a new trend of technologies is emerging whose goal is to elicit user preference in an unobtrusive fashion, for example by visual processing of the facial expressions or audio processing of recorded speech in order to interpret the mood of the user or the analysis of physiological signals such as skin temperature [163] or heart beat [116] among other others. Multimedia processing techniques are required for processing of the data and to transform some these data into useful information for recommendation sources.

[336] proposed a novel type of personalized document recommendation (image, video or text) which replaces target user's ratings with her attention time (AT) over online materials in order to identify/measure her interest over documents. If user's attention time on a target item I_{t_1} is more than item I_{t_2} that is $AT(I_{t_1}, U) > AT(I_{t_2}, U)$ it is reasonable to infer that I_{t_1} is more interesting to user U than I_{t_2}. The key characteristic of the proposed recommendation algorithm is on its ability to track a user's attention time over online materials during his interaction with the online document (reading, browsing or video watching sessions) by employing a vision-based eye-tracking system. Once user's AT over a collection of online documents is obtained, the algorithm predicts the user's probable attention time over a new online item by using data mining of the eye-tracking data.

They key data processing of the proposed method is how to transform the detected fixation points on the screen representing user's gaze into useful data for recommendation. The authors propose three different methods for texts, images and videos in order to anchor these gaze samples onto the corresponding object segments. User attention prediction method for documents is based on word attention using inhomogeneous 2D-Gaussian, image region attention using homogeneous 2D-Gaussian for images and keyframe attention using linear division. The authors apply their proposed technique for similar tasks such as personalized Webpage ranking and user-oriented document summarization based on user attention time [337, 338]. The evaluation is done via a user study by comparing a user-generated rank list of items (reflecting her interests *w.r.t* a given query) as well as the one returned by Google (Web) Search (for texts), Google Image Search (for images) and YouTube (for videos) with those produced by the proposed algorithm. As stated by [223], using eye-movements to build individual models that analyze user behavior patterns for building her profile has been applied in other related tasks such as document filtering [56] and query expansion [57] but their adoption in RS remains vastly under-explored. Conventional ratings given to an item are "atomic", *i.e.,* the user gives her overall preference for the item. One of the main achievement of this type of work in RS is that the preference obtained by user attention is "compositive" which means user gives her rating to subcomponents or different features separately (*e.g.,* for image, user's attention time is the accumulated version of attention over different regions). The main limitation of this research work is that the used set of visual features are old and deprecated. Also, little detail has been shared with the eye gaze tracking module, limiting the reproducibility of the work in practice.

CHAPTER 3

Visual Content-Based Video Recommendation

3.1 Introduction

Video recommender system (VRS) are traditionally powered by either collaborative filtering or by content-based filtering (CBF) engines which are presumably the two most frequently adapted variants of RS today [190, 214, 268]. While collaborative filtering (CF) assumes that the target user (to whom the recommendation is going to be delivered) will like content similar to the content other like-minded users prefer, CB approaches determine their recommendation based on the similarity of the items to the items the user liked in the past. For this, CB approaches rely on descriptive attributes (features) of the items in order to find such similarity, for example colors in an image, motion trajectory in a video, rhythm of a song, textual information of a page.

To the extent CB and their respective algorithm base their similarity computation on either textual metadata or audio-visual features strongly varies between domains, *i.e.,* the type of items recommended. In this regard, while extracting descriptive item features from multimedia content is a well-stabilized research task in the multimedia community, the recommender system community has for long time considered metadata (either *editorial* as in the case of genre, cast, director or *the wisdom of the crowed e.g.,* tags, reviews) as the single source for CB recommendation models, thereby disregarding or to say the least not fully exploiting the wealth of information contained in the actual signals. While such *post-release metadata* are assumed to cover the 'semantics' of the content given that they are generated by human, they have several shortcomings as well:

- **First:** metadata are prone to errors (in particular being user-generated) and labor-intensive/expensive (in particular the editorial ones).
- **Second:** metadata can be rare or unavailable for new videos, making it difficult or even impossible to provide good quality recommendation, *i.e.,* the *cold-start problem.*
- **Third:** even if metadata are available in great amount, they often require complex Natural Language Processing (NLP) in order to account for stemming, stop words removal, synonyms detection and other semantic analysis tasks [19].
- **Fourth:** user-generated metadata are often biased toward users/communities and might not fully or only in a distorted way be able to represent the characteristic of a video [61, 197].
- **Fifth:** the perception of a movie in the eyes of a viewer is influenced and can be manipulated by many film-making conventions applied by the director, *i.e.,* the design aspects of a movie production used to classify aesthetics and style or the so called *mise-en-scène* characteristics of a movie. Although the viewer may not consciously notice, the mise-en-scene elements affect her experience. For example, two movies from the same genre and director "the Empire of the sun" and "Schindler's list" (both drama movies by Spielberg) can be significantly different based on the mise-en-scene characteristics, in which "the Empire of the sun" is shot using bright colors and making heavy use of special effects while "Schindler's list" is shot like a documentary in black and white. Although, these two movies are comparable with respect to traditional metadata, their style and mise-en-scène are substantially different which can likely affect the viewers' feelings and opinions differently.

Addressing these research limitations, we focus on the domain of video recommendation and propose a new type of CB technique that filters videos according to the visual characteristics including lighting, color and motion. The proposed features have a stylistic nature and are in accordance with applied media aesthetics [344] used to convey communication effects and to simulate different feelings in the viewers. Our research hypothesis is that "a RS using low-level visual features (mise-en-scène) provides comparable or better accuracy compared to the same RS using traditional metadata content (genre and tag)". We articulate the research hypothesis along the following research questions:

- **RQ1:** Can the introduction of low-level visual features extracted from videos improve the quality of recommendations? (study 1)

- **RQ2:** Can the introduction of low-level visual features alone or combined with high-level semantic features, induce measurable effects on the *perceived* utility of recommendations? (study 2)

- **RQ3:** Can *hybridization* of the proposed CB visual system with other (traditional) form of recommendation such as matadata-driven CB system or collaborative knowledge shared in a CF system under state of the art hybridization techniques such as factorization machines (FM) or canonical correlation analysis (CCA) improve the quality of recommendation? (study 3)

In this chapter, I seek to answer the above research questions in different sub-studies. Each study in this chapter is the result of one or several publications published at an international conference or peer-reviewed journal.

I believe this chapter provides a number of contributions to the field of RS field in the video domain: first, it improves our understanding of the role of visual information encapsulated in the videos to build automated RS that are capable of responding to the preferences of users, either *autonomously* to replace traditional CB approaches especially in cold-start scenarios or *in conjunction* with other CB techniques based on metadata or the shared community knowledge in CF in order to improve the accuracy of recommendations. To the best of our knowledge, this subject has been heavily under-researched in the RS community; second, as an additional contribution, this study is performed with respect to two entirely different movie datasets (short version: movie trailers v.s. long version: full movies) and provides useful insights from the similarities/difference obtained while extracting the visual features from the complete version of movies or the summarized short version in building CBRS.

3.2 Study 1: "The impact of low-level visual features"

The results of this research study was published at the following venues:

- As a long paper at the *Springer Ec-Web 2015* conference [93]
- As a short paper at the *ACM CHIitaly 2016* conference [94]
- A detailed version of the work involving the study on full-length movies was published at the *Springer Journal of Data Semantics* in 2016 [90].

3.2.1 Background and Related Work

Movie recommender systems typically exploit high-level movie attributes in order to generate movie recommendation [59]. Metadata play an important role in such systems where recommendations are generated using explicit or implicit preferences of users on attributes such as movie genre, director, cast, (structured information) or plot, tags and textual reviews (unstructured information). In contrast, our work exploits "implicit" content characteristics of items, *i.e.,* features that are "encapsulated" in the items and must be computationally "extracted" from them.

We focus on the domain of video recommendations and propose a novel approach for recommendation. The proposed system in study 1 is a visual CB system that filters items according to stylistic visual properties of the videos. We propose a set of computational visual features that can be automatically extracted from video files, either movie trailers or the corresponding full-length videos. Such features include lighting, color, and motion and are referred to as the mise-en-scène throughout this research since they have a "stylistic" nature and are according to applied media aesthetics [344].

A specific interest for this study is applied media aesthetic which is concerned with the connection of a number of media components, such as colors, light, camera movements with the perceptual responses they can bring about in consumers of media communication, mainly videos and films. Such media components, that together form the visual images composing the media are investigated by using a rather formalistic approach that suits the purposes of this study.

By an investigation of cameras, lenses, lighting, and so forth., as production apparatuses as well as their aesthetic characteristics and utilizations, applied media aesthetic attempts to recognize patterns in how such components work to deliver the desired impact in conveying feelings and meanings. The image components that are typically addressed as principal in the literature [233] include lights and shadows, colors, space representation, motion and sound. It has been demonstrated, *e.g.*, in [54, 257], that some aspects concerning these components can be computed from the video information streams as statistical values. We will investigate the details of the features, investigated for content-based video recommendation in this study in order to provide a solid overview on how they are used to producing perceptual responses in the audience. Sound will not be further discussed, since it is out of scope of this study (we will study that in chapter 4), as well as the space representation, that concerns, *e.g.,* the different shooting angles that can be utilized to represent dramatically an event.

The proposed recommendation technique has been evaluated over a large set of video items and the results were compared with existing CB techniques that exploit explicit features such as movie genre. We consider three different experimental conditions: (a) visual features extracted from movie trailers, (b) visual features extracted by full-length video and (c) traditional explicit features based on genre. The goal is to test two hypotheses:

- The recommendation algorithm based on visual features leads to a higher recommendation quality (*w.r.t.* accuracy metrics - See Ch. 5) in comparison with traditional genre-based RS.

- Recommendation quality when stylistic features are extracted from either movie trailers is comparable when they are extracted from full-length videos. In other words, movie trailers are good representatives of their corresponding full-length movies with respected to the utility of recommendation.

The evaluation study has confirmed both hypotheses and has shown that our technique leads to more accurate recommendations than the baselines techniques in both experimental conditions.

3.2.2 Method Description

In the following we will describe the method adopted to build our proposed CB visual recommender system.

Visual feature extraction

The main processing steps involved are: (1) video segmentation and, (2) visual feature extraction. We will describe each of these methods in the following. A video can be considered as contiguous sequence of many frames where consecutive frames are visual highly similar. Considering all these frames for feature extraction is inefficient. Therefore, the first step prior to feature extraction is to segment the video into smaller segments known as *video shots*. A shot boundary is a frame where frames around it have significant difference in their visual content. Frames within a shot have a high-level similarity in their visual appearance, therefore it makes sense to take one representative frame in each shot and use that frame for feature extraction known as the *Key Frame*. Two types of features are extracted from videos: (i) temporal features (ii) spatial features. The temporal features reflect the dynamic perspectives in a video such as the average shot duration (or shot length) and object motion, whereas the spatial features illustrate static properties such as color, light, and so on.

In order to demonstrate the effectiveness of the proposed content, we selected and extracted the five type of visual features proven to be highly representative of movie genres, to be extracted from each video

$$f_v = (L_{sh}, \mu_{cv}, \mu_m, \mu^2_{\sigma_m}, \mu_{lk}) \tag{3.1}$$

where L_{sh} is the average shot length, μ_{cv} is the mean color variance over key frames, μ_m and $\mu^2_{\sigma_m}$ are the mean motion average and standard deviation across all frames respectively and $\mu^2_{\sigma_m}$ is the mean lightening key over key frames.

Shot Duration: The shot duration measures how long a shot is onscreen before transitioning to a new shot [97]. A shot is a single camera action, and the total number of shots in a video is indicative of the *pace* at which the movie is being created. For instance, the genre action typically contain many camera movements (to follow people and objects in the environment) in comparison with drama genre which usually contain long conversation scenes as shown in Figure 3.1. The speed of cut sequences influences perceptual and emotional elements in the viewer [90, 97, 257]. The average shot duration is defined as

$$\overline{L} = \frac{n_f}{n_{sh}} \tag{3.2}$$

where n_f is the number of frames and n_{sh} the number of shots in the movie.

Color Variance: Colors can influence our feelings in different ways. They can bring excitement and joy, makes us more aware of the world around us [344]. Variance of colors in movies is highly correlated with the style of the movie [257, 299]. For instance, filmmakers often use a large variety of light colors for shooting comedy movies and

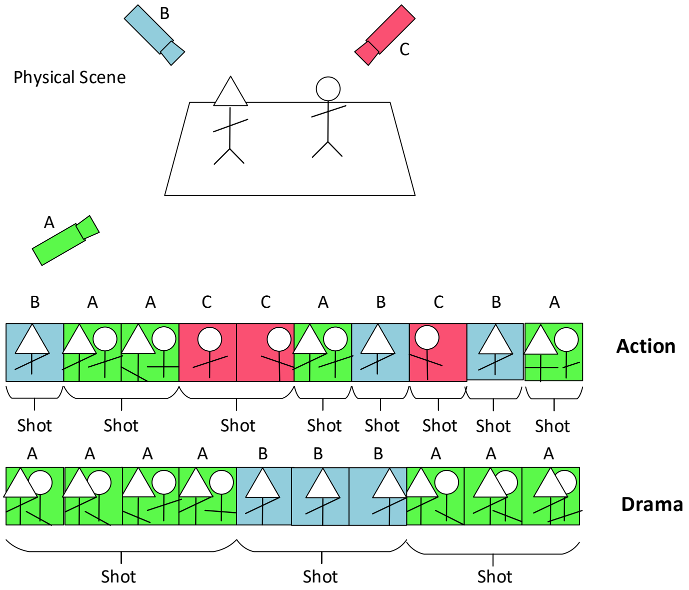


Figure 3.1: *Illustration of camera shot change speed in different movie genres. The genre action typically contains high number of shots (i.e., high activity) with short duration as opposed to drama genre with less number of shots and long scene duration.*

less combination for horror movies. For each key frame represented in Luv color space [274], we first compute the covariance matrix ρ as

$$\rho = \begin{pmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{pmatrix} \tag{3.3}$$

where σ_L, σ_u, σ_v, σ_{Lu}, σ_{uv}, σ_{Lv} are the standard deviation and mutual covariance over three channels L, u, v. We later compute the generalized color variance γ as the determinant of ρ

$$\gamma = \det(\rho) \tag{3.4}$$

We use γ as the output representation of the color variance within a shot [257].

Average and Variation of Motion: Motion is the *basic structural unit* of video and it is what differentiates it from a static image. Motion in a video can be due to camera movements (camera motion) or movements of objects in the environment (object motion). For each keyframe within a shot, with respect to its previous frame, we estimate the velocity (motion) of each pixel by using the optical flow technique [153]. This results in a motion vector calculated at each pixel. The average motion magnitude contains information about the amount of activity in that frame. In order to capture both motion types we computed the average (mean) and variation (variance) of motion vectors magnitudes.

Lighting: Lighting is the deliberate manipulation of light and shadows to communicate a specific purpose. Certain colors and lighting can have an immediate emotional effect on the consumers as they can establish the aesthetic context for our experiences, a framework that tells us how we should feel about a certain event. These elements sidestep our rational faculties and are used in the hands of able filmmakers for controlling the type of emotions intended to be induced to a movie consumer [344]. There are two main lighting categories: *high-key lightening* and *low-key lightening*. With high-key lightening there is abundance of light and less contrast between dark and light, as with comedy movies. With low-key lightening there is predominance of darker tones and a high contrast ratio, as with noir films. In order to measure lightening, for each key-frame we first compute mean μ and standard deviation σ of the brightness of each pixel. Lighting key ξ for each keyframe is defined as

$$\xi = \mu \cdot \sigma \tag{3.5}$$

Recommendation

We formulate the recommendation problem as follows. The unknown preference score (*i.e.,* rating) $\hat{\mathbf{r}}_{ui}$ for user u and item i is computed as an aggregate of the ratings of other, similar items, using the following aggregation function:

$$\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{j \in N_u(i)} s_{ij}} \sum_{j \in N_u(i)} s_{ij}\, \mathbf{r}_{ui} \tag{3.6}$$

where $N_u(i)$ denotes the items rated by user u most similar to item i and s_{ij} is the similarity score between items i and j (the content-centric similarity). Also known as *matrix completion* problem, recommendation is realized by rating prediction for various user-item combinations and then ranking the predictions. We searched for best number of neighbors in the range $[2 - 10]$ for each recommender and reported the final results under the best parameter setting.

3.2.3 Evaluation Methodology

In order to test the research hypothesis presented in section 3.1, we evaluated the Top-N recommendation quality of each CBRS on a subset of the MovieLens-20M dataset [142] by employing 5-fold cross validation (CV) in our experiments on (rating splitting). The evaluation is conducted by computing Recall@K defined as:

$$R@K = \frac{1}{|U|}\sum_{u\in U}\frac{|L_u \cap \hat{L}_u|}{|L_u|} \tag{3.7}$$

where L_u is a set of relevant items of user u in the test set T and $\hat{L}_u$ denotes the recommended set containing the K items in T with the highest predicted ratings for the user u from the set of all users U. We have selected a subset of full-length movies and the corresponding trailers, that were sampled randomly from all the main genres, *i.e.,* Action, Comedy, Drama and Horror. The summary of the final dataset is given in Table 3.1.

Table 3.1: *General information about our dataset*

# items	**167**
# users	139190
# ratings	570816

As noted before, the movie titles were selected randomly from MovieLens dataset, and the files were obtained from *YouTube*[1]. The dataset contained over all 167 movies, 105 of which belonging to a single genre and 62 movies belonging to multiple genres (see Table 3.2).

Table 3.2: *Distribution of movies in our catalog*

	Action	Comedy	Drama	Horror	Mixed	**Total**
#	29	27	25	24	62	**167**
%	17%	16%	15%	14%	38%	**100%**

The proposed video feature extraction algorithm was implemented in MATLAB R2015b [2] on a workstation with an Intel Xeon(R) eight-core 3.50 GHz processor and 32 GB RAM. The Image Processing Toolbox (IPT) and Computer Vision Toolbox (CVT) in MATLAB provide the basic elements for feature extraction and were used in our work for video content analysis.

[1] www.youtube.com

[2] http://www.mathworks.com/products/matlab

In addition, we used the R statistical computing language [3] together with MATLAB for data analysis. For video classification, we took advantage of many of the classifiers in Weka [4] that provides an easy-to-use and standard framework for testing different classification algorithms.

3.2.4 Results

Genre Classification

In this section, the automated classification of movie trailers into genre by using the mise-en-scène visual features is studied where a single movie may belong to more than one class. The target studied genres are: *Action*, *Comedy*, *Drama* and *Horror* as commonly practiced in movie domain [257, 299, 351]. Being able to classify movies into genres allows the proposed CB visual system to be safely and better replaced with the genre-based RS which can be advantageous useful in cold-start situations where no or little metadata are available.

We assume that an item is represented by a feature vector of d attribute values $x = [x_1, ..., x_d]$ and there exists a set of m items $I = \{i_1, i_2, ..., i_m\}$ where each item (instance) can be associated with a subset of labels from $\mathcal{L} = \{1, ..., L\}$. The goal in multi-label classification is to model a classifier $\mathbf{h}$ able to associate a set of c labels to every item in I, where $c \in [1, L]$ varies for every item. This is opposed to the traditional task of single-label classification (*i.e.* multi-class, or binary) where each instance is only associated with a single class label. A common approach to multi-label classification is to perform *problem transformation*, whereby a multi-label problem is transformed into one or more single-label (*i.e.,* binary, or multi-class) problems. In this way, single-label classifiers are employed and their single-label predictions are transformed into multi-label predictions. Further information can be found in [347].

In this experiment, since the goal if to focous on the recommendation quality of the proposed visual CB system and to a lesser extent on the classification part, we made a simplifying choice for the genre classification task and considered in our original dataset, only movies tagged with one single genre label, thereby effectively changing the original problem from a multi-label classification problem to a binary-classification problem.

[3] https://www.r-project.org

[4] http://www.cs.waikato.ac.nz/ml/weka

Experiment A

For the classification task, we considered 105 movies (from the original 167) tagged with one of the four main genres (action, comedy, drama, horror) as suggested in [257]. We experimented with many classification algorithms in Weka and obtained the best results under *decision tables* [184]. Decision tables can be considered as tabular knowledge representations [185] which given a new instance it searches for an exact match in the decision table cells, and then the instance is assigned to the most frequent class among all instances matching that table cell [121].

The following classification quality *w.r.t* accuracy was obtained after running a 10-fold cross-validation: (1) movie trailers: **76.2%** and (2) full-length movies: **70.5%**. The best and worst classification accuracy occurred for the genre comedy (23 out of 27 correct) and the genre horror. As foe the latter, for instance, 4 out of 24 horror movie trailers have been mistakenly classified as action genre. This phenomenon can be explained by the fact that there are many action scenes occurred in horror movies, and this may make the classification very hard. Similar trends of results were observed for full-length movies. From the results obtained in this experiment it can be concluded that the low-level stylistic visual features used in our experiment are predictive of genre classes by a proper extent.

Correlation between Full-length Movies and Trailers

As mentioned earlier in section 3.1, one of the research hypotheses we are seeking for in this research is if there is correlation between the full-length movies and their corresponding trailers given the mise-en-scène visual features extracted from each dataset. The goal is to investigate whether or not the trailers are representative of their full-length movies, with respect to the stylistic visual features.

Such a correlation can be defined in different ways: (1) by directly computing pairwise correlation between low-level visual features and, (2) by computing the recommendation quality *w.r.t.* a certain quality metric using each of the two movie datasets. In the light of above, we have performed two experiments as explained in the following.

Experiment B

We have first extracted the five low-level visual features defined in section 3.2.2 from each of the 167 movies and their corresponding trailers in our dataset. Similarity between the two datasets is computed by calculating the pairwise correlation using the cosine similarity metric and computing

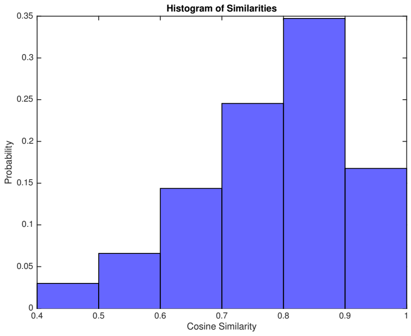


Figure 3.2: *Histogram distribution of the cosine similarity between full-length movies and trailers*

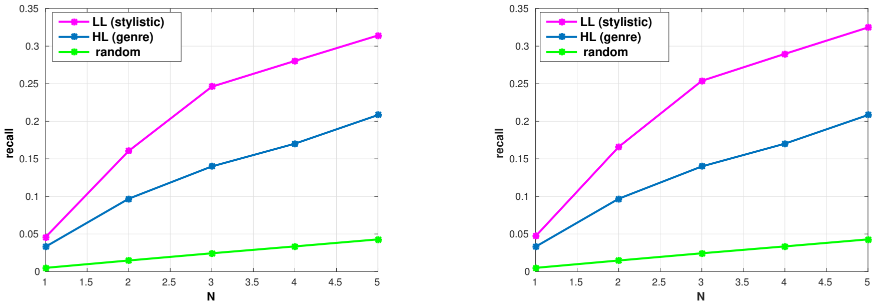
the average across all items. The reason for the choice of cosine metric is that it is the the same metric used to generate recommendations in our experiments (as explained in Section 3.2.3), and hence, it is a reliable indicator to evaluate if recommendations based on trailers are similar to recommendations based on the full-length movies.

Figure 3.2 plots the histogram of the cosine similarity. As it can be seen, the average similarity between the two dataset is approximately **0.78** and the median is **0.80**. It can be also noted that more than **75%** of the movies have a cosine similarity greater than 0.7 between the full-length movie and trailer. The movies most similar to their trailer are “Evil - In the Time of Heroes”, “Munger Road”, and “Pacific Rim” while the least correlated movies are “Love is Strange”, “The Resident”, and “Die Hard: With a Vengeance”. Moreover, less than 3% of the movies have a similarity below 0.5.

Overall, the cosine similarity shows a substantial level of correlation between the movie trailers and full-length videos. This is an interesting outcome that basically indicates that the trailers of the movies can be considered as good representatives of the corresponding full-length movies.

Experiment C

In the second experiment in study 2, we have used the low-level features extracted from both trailers and the full-length movies to feed the CBRS described in Section 3.2.2 in order to evaluate if there is significant difference between the quality of recommendation obtained from the two datasets. In absence of such a significant difference, one can reliably replace the shorter version of the movies for conducting research in the field of RS. Quality



(a)

(b)

Figure 3.3: *Performance comparison of different CB methods under best feature combination for full-length movies (a), and trailers (b).*

of recommendations has been evaluated according to the methodology described in Section 3.2.3.

Figure 3.3 plots the recall@N for full-length movies (a) and trailers (b), with values of N ranging from 1 to 5. Recall reported in the figures the algorithm in both figures are the values for the similar algorithms, but using different sources of information (*i.e.,* low-level vs high-level content features). The best number of neighbors have been determined with cross validation and were $K = 2$ for LL features and $K = 10$ for HL features).

By comparing the two figures, it is can be noted that the recall values of the CBRS using the visual features extracted from the full-length movies and trailers are almost identical. These results have interesting implication and confirm those obtained in Experiment B (of the same study). The result prove that low-level features extracted from trailers are representative of the corresponding full-length movies and can be effectively used to provide recommendations.

Recommendation Quality compared with metadata

The goal of this section is to investigate the main research hypothesis: "if low-level visual features can be used to provide good-quality recommendations". We compare the quality of CB recommendations based on three different types of features:

Low Level (LL): stylistic visual features. (aka mise-en-scène features)

High Level (HL): semantic features based on genres.

Hybrid (LL+HL): combination of stylistic and semantic features.

Experiment D

In order to identify the visual features that are more useful in terms of recommendation quality, we have performed an extensive set of experiments. We have fed the CBRS with with all the 31 combinations of the five visual features f1-f5 (as for LL feature), one genre feature vector (as for HL features) and 31 additional combinations for low-level stylistic visual features with the genre. Each of the resulting feature vectors were max normalized.

Table 3.3 reports *Recall@5* for all the different experimental conditions. The first column of the table describes which combination of low-level features has been used (1 = feature used, 0 = feature not used). The last column of the table reports, as a reference, the recall when using genre only and it does not depend on the low-level features. The optimal value of number of neighbors used for the KNN similarity has been determined with cross validation and shown in each column of the table.

From these results it can be seen that the recommendation quality *w.r.t* recall@5 are clearly better for low-level stylistic visual features extracted from trailers than recommendations based on genre for **any** combination of visual features. However, no considerable difference can be observed between genre-based and hybrid-based recommendations.

3.2.5 Feature Analysis

The goal of this is study to investigate if and how some low-level visual features provide better quality of recommendation. We also wish to improve our understanding why combinations of low-level features do not improve accuracy.

Experiment E

In this experiment, we analyze if there is a correspondence/correlation between the two video datasets: trailer and full-length movies. This analysis is similar to the one reported in Section 3.2.4, but results are reported as a function of the features.

Figure 3.4 plots the cosine similarity values between visual features extracted from the full-length movies and visual features extracted from trailers. In other words, for each of the features extracted, we compute the pairwise cosine similarity between the same feature across all video item. As it can be seen features f2 and f4 (color variance and object motion) are the least similar features, suggesting that their adoption, if extracted from trailers, should provide different recommendation quality for the two datasets.

Table 3.3: *Performance comparison of different CB methods, in terms of Recall metric, for different combination of the Stylistic visual features*

Features					**Recall@5**		
					LL Stylistic	**LL+HL Hybrid**	**HL Genre**
f_1	f_2	f_3	f_4	f_5	$(K = 2)$	$(K = 2)$	$(K = 10)$
0	0	0	0	1	0.31	0.29	0.21
0	0	0	1	0	0 32	0.29	
0	0	1	0	0	0.31	0.22	
0	1	0	0	0	0.27	0.23	
1	0	0	0	0	0.32	0.25	
0	0	0	1	1	0.32	0.21	
0	0	1	0	1	0.31	0.22	
0	0	1	1	0	0.32	0.22	
0	0	1	1	1	0.32	0.23	
0	1	0	0	1	0.24	0.20	
0	1	0	1	0	0.25	0.20	
0	1	0	1	1	0.25	0.22	
0	1	1	0	0	0.24	0.20	
0	1	1	0	1	0.23	0.20	
0	1	1	1	0	0.25	0.18	
0	1	1	1	1	0.25	0.22	
1	0	0	0	1	0.31	0.26	
1	0	0	1	0	0.31	0.29	
1	0	0	1	1	0.31	0.18	
1	0	1	0	0	0.30	0.23	
1	0	1	0	1	0.30	0.22	
1	0	1	1	0	0.31	0.24	
1	0	1	1	1	0.31	0.23	
1	1	0	0	0	0.25	0.20	
1	1	0	0	1	0.23	0.20	
1	1	0	1	0	0.25	0.22	
1	1	0	1	1	0.25	0.21	
1	1	1	0	0	0.22	0.20	
1	1	1	0	1	0.21	0.20	
1	1	1	1	0	0.25	0.21	
1	1	1	1	1	0.25	0.21	

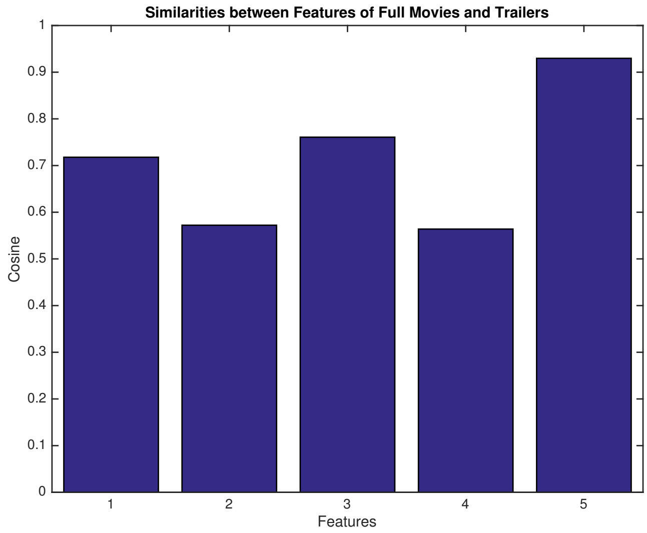


Figure 3.4: *Cosine similarity between stylistic visual features extracted from the full-length movies and their corresponding trailers*

A significance hypothesis test using the Wilcoxon test was adopted to compare features extracted from the full-length movies and trailers and the results are summarized in Table 3.4. The results of the significance test indicate that there is a significance difference between features f1 (average shot length), f2 (color variance) and f4 (motion variation) while no significance difference is obtained for f3 (motion average) and f5 (lighting key). These results show that movie trailer and the correspondence full-length videos are highly similar *w.r.t.* features f3 (motion average) and f5 (lighting key) and that the former can safely represent the latter dataset for example when full-length videos are not accessible.

Table 3.4: *Significance test with respect to features in 2 set of datasets (movie trailers and full movies)*

	$f_1(\overline{L}_{sh})$	$f_2(\mu_{cv})$	$f_3\ (\mu_{\overline{m}})$	$f_4\ (\mu_{\sigma_m^2})$	$f_5(\mu_{lk})$
wilcox.test	1.3e-9	5.2e-5	**0.154**	2.2e-16	**0.218**

Having considered all these results, we remark that our considered hypotheses have been successfully validated. In fact, our results indicate that a proper extraction of the visual stylistic features of videos can improve the accuracy of video recommendation in comparison with typical expert annotation method, for both situations where the visual features are extracted from full-length videos or from movie trailers only. These are promising results, as they overall illustrate the possibility to achieve high-accurate recommendation qualities with an automatic method than a manual method (*i.e.,* expert annotation of videos) where the manual method can be very costly and in some cases even infeasible to obtain (*e.g.,* in presence of large datasets).

3.2.6 Discussion

Our main research hypothesis is that low-level artistic (stylistic) features (*e.g.,* colors, light, motion) can be more representative than high-level semantic features (*e.g.,* genre) in providing content-based recommendations. Our secondary research hypothesis is that low-level features extracted from a movie trailer are representative of the low level features extracted from the full-length movie. We discuss each of these research questions in the following:

Quality of Recommendations

According to the results presented in Table 3.3, all combinations of the low-level visual features provide better recommendation than the high-level feature (genre). The improvement is particularly evident when using either scene duration, light, camera movement, or object movement, with an improvement of almost 50% in terms of recall with respect to genre-based recommendations. The improvement is less strong when using color variance, suggesting that user opinions on average are less affected by how diverse colors are used in movies. We also think the validity of these findings can be restricted to the actual experimental conditions considered, and may be affected by the limited size of the dataset or user-centric factors such as user age. Despite these limitations, our results provide empirical evidence that

> *A set of low-level visual features based on mise-en-scène obtained from applied media aesthetic [90, 93, 344] can provide predictive power, comparable to the genre of the movies, in predicting the relevance of movies for users.*

Surprisingly, mixing low-level and high-level features does not improve the quality of recommendations and, in most cases, the quality is reduced with respect to use of low-level only, as shown in Table 3.3. This can be explained by observing that genres can be easily predicted by low-level features. For instance, action movies have shorter scenes and shot lengths, than other movies. Therefore, in presence of correlation between features, the overall prediction capabilities of the mixed approach is reduced.

Trailers vs. Movies

One of the potential drawbacks in using low-level visual features is the computational load required for the extraction of features from full-length movies. In some cases these movies may be inaccessible or costly to obtain specially for conducting a research study using a large number of videos. The results of our research indicates that low-level features extracted from movie trailers are strongly correlated with the corresponding features extracted from full-length movies (average cosine similarity **0.78**) in which scene duration, camera motion and light are the most similar features when comparing trailers with full length movies. The result for the scene duration is somehow surprising, as we would expect scenes in trailers to be, on average, shorter than scenes in the corresponding full movies since movie trailers are usually made with many intentional/abrupt cuts. The strong correlation however suggests that trailers have *consistently* shorter shots than full movies. For instance, if an action movie has, on average, shorter scenes than a dramatic movie, the same applies to their trailers. Our results provide empirical evidence that

> *The tested set of low-level visual features based on mise-en-scène extracted from trailers can be used as an alternative to features extracted from full-length movies in building content-based recommender systems.*

3.2.7 Conclusion of Study 1

In study 1, we presented a novel CB method for the video recommendation task. The method extracts and uses the low-level visual features from video content in order to provide users with personalized recommendations, without relying on any metadata features such as genre, cast, reviews - which are more costly to collect, because they require an "editorial" effort, and are not available in many new item scenarios.

We have developed a main research hypothesis, *i.e.,* a proper extraction of low-level visual features from videos may led to higher accuracy of video recommendations than the typical expert annotation method. Based on a large number of experiments, we have successfully verified the hypothesis showing that the recommendation accuracy is higher when using the considered low-level visual features than when high-level genre data are employed. The findings of our study do not diminish the importance of explicit semantic features (such as genre, cast, director, tags) in content-based recommender systems. Still, our results provide a powerful argument for exploring more systematically the role of low-level features automatically extracted from video content and for exploring them.

3.3 Study 2: "User study"

The results of this research study was published as a short paper at the *ACM Recsys 2017* conference [112].

3.3.1 Introduction and Context

As the world wide web has become the main source and distribution channel of digital videos and movies, a large amount of videos are accessible to users. Video sharing websites, such as YouTube, Netflix and Hulu, host a tremendous number of videos. These massive video repositories place an enormous burden on users when trying to find videos of interest. In order to curb this *multimedia information overload* and allow users to find their desired video content, most video websites have adopted recommender systems, as an effective way to help users explore the world of videos [23,268]. Successful recommender systems produce well-fitting yet novel and diverse recommendations.

In the last decade, we have seen much consideration given to the *semantic gap problem* in multimedia information retrieval (MMIR) systems [166, 217]. This problem refers to the gap between the high-level concepts that users expect when searching for interesting multimedia content when searching for an item explicitly via a query (*e.g.*, genre, plot, actors) and the low-level features that it is conceivable to automatically extract from the content (*e.g.*, brightness, contrast, *etc.*). The users' ability to effectively and efficiently use MMIR systems is affected by the sharp discontinuity that exists between the crude low-level features and the (semantically) richness of user queries encountered in multimedia search. As a consequence, the MMIR community has since a long time ago struggled to *bridge* this semantic gap as various studies have acknowledged the difficulty of addressing the user's information needs with primitive low-level features, since they are not concerned with the semantics of the content [105, 166].

Despite this fact, we make a different assumption for recommender systems. We still believe that the semantic gap exists and that adoption of low-level features causes limitations in MMIR applications where the objective is to provide a manner for indexing multimedia content so that users can *explicitly* (*i.e.*, manually) query that content at the semantic level. We wish to explore if this assumptions holds additionally for RS, where the objective is to *automatically* discover content that the user likes, without the need for the user to ask the system through querying. Our research assumption is that the semantic gap is not a problem for RS but an opportunity. Our goal is to leverage the low-level features naturally available in the multimedia content and complement them with the high-level characteristics of the content provided by the wisdom of the crowd. For this purpose, we investigate and assess a recent approach for movie recommendations that integrates traditional high-level semantic attributes, such as genre, director and cast, with low-level mise-en-scène features, *i.e.*, the design aspects of movie making influenced by aesthetic and style [90, 344]. Examples of mise-en-scène characteristics include but not limited to lighting, colors, background, and movements.

We believe that mise-en-scène features provide the chance to make video RS more effective and helpful as they help in strengthening two weak spots when working with semantic attributes: *absence of diversity* and *novelty* in video recommendations [120]. Previous research works provide evidence that the lack of diversity and novelty in conventional RS occurs because recommender algorithms are designed to recommend videos similar to the ones users liked in the past [60, 120, 228]. Recommendations lacking novelty and diversity have negative outcomes on user satisfaction, even if recommendations superbly match users' tastes [62, 345]. This paper expands on previous research studies identifying mise-en-scène characteristics that conceivably influence accuracy of recommendations from system-centric viewpoint. In this work we aim to address the following research questions:

RQ1: Can the introduction of low-level visual mise-en-scène features based on mise-en-scène characteristics combined with high-level semantic attributes, improve *offline quality* of video recommendations?

RQ2: Can the introduction of low-level visual mise-en-scène features based on mise-en-scène characteristics combined with high-level semantic attributes, induce measurable impacts on the *perceived utility* of recommendations?

The two research questions presented have been explored by conducting two wide and articulated experimental studies: (a) a system-centric evaluation to measure the offline quality of recommendations in terms of precision, novelty, diversity and coverage; (b) user-centric online experiment involving 100 users, measuring different subjective metrics (*i.e.,* relevance, novelty, diversity, and satisfaction). In both studies, the quality of recommendations have been evaluated under three different experimental conditions characterized by the same CB algorithm using either (i) semantic high-level movie attributes (ii) mise-en-scène low-level features, and (iii) a combination of the two as summarized in Table 3.5.

Results of the study indicate that (1) the adoption of mise-en-scène features can strongly affect precision/relevance, diversity and novelty of recommendations, determining an increased utility of recommendations and in influencing the user's choice to actually watch a movie, and (2) the introduction of mise-en-scène features in conjunction with traditional attributes has a tenancy to diversify recommendations and suggest users with less evident choices.

3.3.2 Experiment Setup

Table 3.5 summarizes the experimental conditions used in two studies.

Recommendation Algorithm

In order to generate recommendations, in both offline and online experiments, we used a widely used pure CF algorithm based on k-nearest neighbors, and considered 20 neighbors, cosine similarity, and log-quantile normalization of the mise-en-scène features [90].

Content-based Features

In both studies, the quality of recommendations have been evaluated under three different experimental conditions defined by one manipulatable variable: the type of movie features.

Metadata features.

Content-based movie recommender systems are conventionally based on high-level attributes such as genre, director, and actors, the so called metadata. Such metadata are generated by human, either editorially (*e.g.* title, cast) or by leveraging the wisdom of the crowd (*e.g.*, tags).

Mise-en-Scène features.

As shown in study 1 [90], low-level stylistic characteristics – such as color, motion and lighting – have been shown effective for CB video recommendations (mise-en-scène features). Mise-en-Scène features can be extracted automatically by processing the video files. Other types of visual features

Table 3.5: *Experimental conditions used in two studies*

Study	Feature Types	Quality Metrics	#Users
A: Offline	(i) semantic (ii) mise-en-scène (iii) combination of semantic and mise-en-scène	Precision Diversity Novelty Coverage	113,682
B: Online		Relevance Diversity Novelty Satisfaction	100

explored in the community of multimedia recommender systems can be found in [146, 168, 270]. The common approach to extract this five categories of features comprise of the following steps:

- videos are partitioned into (non-overlapping) shots by using a shot-boundary detection method. Then, for each shot, a representative *key frame* is extracted (this is not necessary for the shot duration);
- for each shot, the desired features are extracted.
- the feature values are averaged over all shots.

In our experiments, we extracted the five categories of mise-en-scène low-level features described in study 1 and the research works [90, 257] as they are explainable, easy to extract from video files, and show promising results in offline experiments: *shot duration*, *color variance*, *lighting key*, *average motion*, and *motion variation*.

3.3.3 Study A: Offline Experiment

In the offline experiment, we evaluated the Top-N recommendation quality of each CB algorithm powered by one out of three type of content feature presented in Table 3.5. The evaluation of the performance of the RS was performed using a subset of the MovieLens-20M dataset [142], by running a hold-out (80% train - 20% test). The characteristics of the final dataset is shown in Table 5.3:

As for semantic attributes, each video (movie) is described by a feature vector of length 127 based on user-generated tags where on average each movie is labeled with six attributes. This is while the mise-en-scène dataset contains five feature values extracted according to the procedure

Table 3.6: *Characteristics of the evaluation dataset used in the offline study: $|\mathcal{U}|$ — number of users, $|\mathcal{I}|$ — number of items, $|\mathcal{R}|$ — number of ratings.*

dataset	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$	$\|\mathcal{R}\|$	$\frac{\|\mathcal{R}\|}{\|\mathcal{U}\|}$	$\frac{\|\mathcal{R}\|}{\|\mathcal{I}\|}$	$\frac{\|\mathcal{R}\|}{\|\mathcal{I}\|\times\|\mathcal{U}\|}$ (density)
ML-20M	113,682	12,573	775,090	6.81	61.6	0.0005

described in Section 3.3.2. The offline quality of recommendation is evaluated *w.r.t.* the following accuracy and beyond-accuracy metrics: *precision*, *diversity*, *novelty* and *coverage* [133].

Precision is computed by calculating the the percentage of movies in the suggestion list that were relevant to the user. Diversity is estimated by computing the intra-list similarity between items *w.r.t.* certain content feature. We computed diversity based on *genre* since it can be a considered as fair mediator between the mise-en-scène features and tag semantic features not biased toward one. Diversity is quantified by calculating the pairwise cosine similarity between items in the recommendation list and subtracting it from one given as $1-S$ where S is the average pair-wise similarity between all the items in the list. Novelty is measured *w.r.t.* popularity of the items using the self-information of the recommended items relative to their global popularity. Lastly, we compute coverage by calculating the percentage of pairs of $< user, movie >$ for which we can predict a rating [22, 110, 126]. After calculating the above, all metric values were normalized to the range 0% to 100%, representing the lowest and highest quality respectively. See section 5.6 for further information on the evaluation metrics.

3.3.4 Study B: Online Evaluation

Perceived Quality Metrics.

The objective of this study is to measure the user's perceived quality of the recommendations. Perceived quality is the extent by which the users judge recommendations fruitful and appreciates the overall experience with the RS. The former is operationalized by measuring the following evaluation metrics: *perceived accuracy*, *novelty*, *diversity*, and overall *user satisfaction* as defined in [110, 182, 254]:

1. *Perceived accuracy* (aka *Relevance*) measures how much the recommendations match the users' interests, preferences and tastes;

2. *Diversity* measures how much users perceive recommendations as different from each other, *e.g.,* movies from different genres;

3. *Novelty* measures the extent to which users receive new recommended movies;

4. *Overall Users' Satisfaction* measures the global users' feeling of the experience with the recommender system.

Procedure.

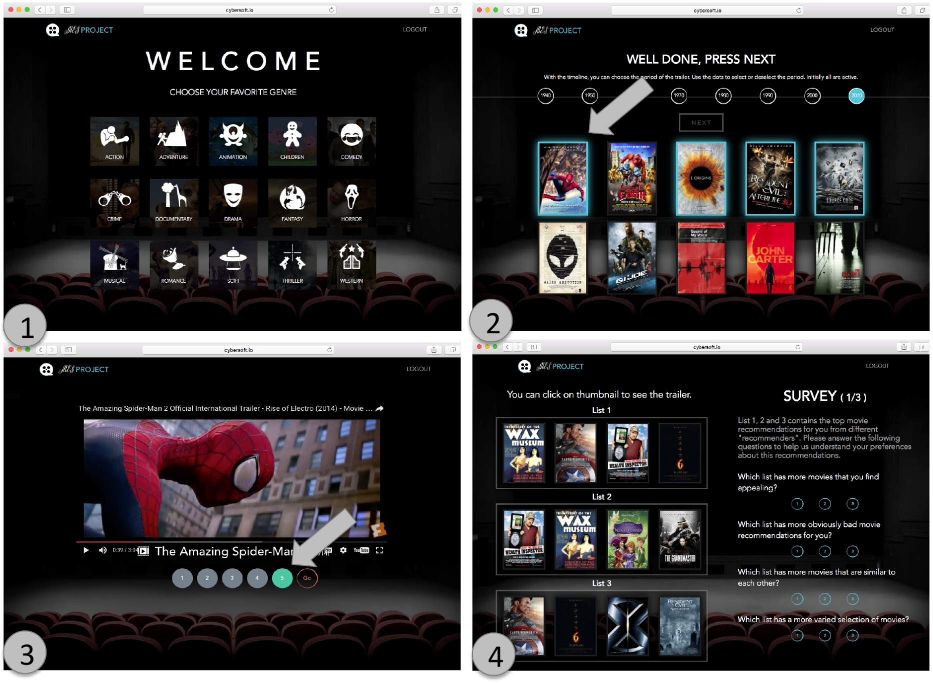
For the purpose of our study, we have developed *MISRec*[5], a a test framework based on web for the movie domain, which can be effortlessly configured to assist the execution of controlled empirical studies. MISRec is powered by the same pure CBF algorithm described in Section 3.3.2 and supports users with a variety of functionalities that are common in online video-streaming services such as Netflix and Lovefilm (Figure 3.5). MISRec contains the same catalog of movies used in the first study in which users can browse a catalog of movies, retrieve detailed description about each movie, rate them, and receive recommendations. MISRec also provides a questionnaire system that permits researchers to collect quantitative and qualitative information from the user in a relatively easy manner.

Our primary research audience is represented by users in the age range between 20 and 50 who have some familiarity with the use of the Web and had never used MISRec before the study. This to control for the potentially mystifying factors of biases or misconceptions derived from previous uses of the system. The total number of recruited subjects who completed the task was 100 (63% men, 37% women, average age: 27.8, std 3.83).

The interaction starts with a sign-up procedure, where each participant (user) is initially requested to provide her basic demographic information. Afterwards, she is invited to browse the movie catalog, and select five movies and rate them using a 5-level Likert scale. (1 = low interest for/appreciation of the movie; 5 = high interest for/appreciation of the movie). This process is formally known as *preference elicitation* [254]. Based on these ratings, three recommendation lists are generated - one for each experimental condition - each list containing 4 recommended movies.

Since it is not practically feasible to ask the user to watch the full-version of movies and given that the user may have already seen some of the recommended movies, in order to take into account for both of these facts, the user is asked to watch/preview the shorter trailers of the recommended movies within each of the 3 lists and reply to a set of questions related to the quality of the recommendations. Users are then asked to indicate their

[5] short for Mise-en-Scène Movie Recommender



With the timeline, you can choose the period of the trailer. Use the dots to select or deselect the period. Initially all are active.

You can click on thumbnail to see the trailer.

List 1, 2 and 3 contains the top movie recommendations for you from different "recommenders". Please answer the following questions to help us understand your preferences about this recommendations.

Which list has more movies that you find appealing?

Which list has more obviously bad movie recommendations for you?

Which list has more movies that are similar to each other?

Which list has a more varied selection of movies?

Figure 3.5: *Example screenshots of the devloped MISRec movie recommender application [112]*

responses to each of the questions by selecting one of the three lists. Recent research suggest effectiveness of responses obtained in a comparative way compared with absolute manner [110]. To avoid possible biases, the positions of the recommendation lists were randomized for each user. It is worth noting that since watching trailers is a slow process, we decided to include 4 recommendations within each of the 3 lists. A subgroup of the questionnaire as reported in [110, 182] was used to measure the perceived quality of recommendations:

- **Acc.** Which list better understand your taste in movies?
- **Div.** Which list has movies that match a wider variety of moods?
- **Novelty.** Which list has more movies that are familiar to you?
- **Satisfaction.** Which list would you be more likely to recommend to your friends?

We also posed a number of equivalent questions expressed in a different way, in order to verify for consistency of answers.

3.3.5 Experiments and Results

In this section we present the results of the two studies.

Experiment A: Offline Evaluation

Table 3.7 presents the offline quality of the recommendations for each of the three experimental conditions:

- *semantic attribute*: traditional high-level semantic attributes of movies using metadata;
- *mise-en-scène*: low-level visual features based on mise-en-scène;
- *hybrid*: a combination of mise-en-scène features and semantic attributes.

As for the last approach, the hybridization can be done by using an early or late fusion approach (aka. feature-level or ensemble-level) [91]. We chose fusion that is more similar to the latter given its ease of implementation. The ensemble-level hybridization in our system was implemented by interleaving the recommendation results based on movie attributes and mise-en-scène features [22]. The quality of recommendation was measured with the offline methodology as described in Section 4.6. Analysis of variance suggests that the three experimental conditions have a significant impact ($p < 0.05$) on the four variables: precision, diversity, novelty, coverage. For each quality metric, statistically significant values are highlighted in bold.

The results for *precision* indicates that recommendations generated solely based on the mise-en-scène features are the least accurate, while both the hybrid approach and the traditional approach are the most accurate. If we look at *diversity* and *coverage* of recommendations, both approaches based on mise-en-scène features (alone or hybridized with semantic attributes) provide the best recommendations, *i.e.*, the most diverse recommendations, able to span almost all of the items in the catalog. As for novelty, it can be noted that the approach based solely on mise-en-scène features provides the best recommendations, with the hybrid approach being marginally better than the approach based on semantic attributes.

Table 3.7: *Results of experiment A: Offline evaluation. Results in bold are significantly different ($p < 0.05$)*

Research Variables	Semantic attributes	Mise-en-Scène features	Hybrid
Precision	**16.9%**	6.3%	**16.1%**
Diversity	20.8%	**21.8%**	**21.9%**
Novelty	94.9%	**96.7%**	95.7%
Coverage	75.5%	**92.7%**	**92.7%**

Experiment B: Real User Study

In this section, we describe an empirical user study that considers the same recommender algorithms used for the objective evaluation in Experiment A, but measures the quality of the recommendations perceived by real users. We have analyzed the opinions and behavior of 100 users interacting with an online and real-time movie recommender system.

We initially cleaned the gathered information by removing the data referring to subjects who indicated clear evidences of gaming with the testing system. We removed the participants who interacted with the system for under 2 minutes, or left some questions unanswered. We also introduced a number of comparable questions formulated in a different way (*e.g.,* negatively), so as to check for irregularity of answers. In the final questionnaire participants were requested to choose the favored recommendation list (binary choice). We performed multiple pair-wise Cochran Q tests on the responses from the users, as it well fits to the characteristics of the gathered data [296] (binary responses of type *v.s.* all) . All tests were run considering significance level $\alpha = 0.05$.

Table 3.8: *Results of the experiment B: Real User Study. Results in bold are significantly different ($p < 0.05$).*

Research Variables	Movie attributes	Mise-en-Scène features	Hybrid
Relevance	25%	15%	**60%**
Diversity	25%	22%	**53%**
Novelty	21%	19%	**60%**
Satisfaction	21%	22%	**57%**

The final results for the online evaluation can be found in Table 3.8. The results indicate that the adoption of mise-en-scène features alone provides the lowest perceived quality *w.r.t.* traditional semantic attributes, although the difference if not significant (for accuracy, diversity, and novelty). These results are partially in contrast with the previous study, in which novelty and diversity with mise-en-scène recommendations were shown significantly better than with traditional attributes. This could be explained by previous works suggesting that offline evaluations metrics are not always good predictors of the perceived quality of recommender systems [77, 188]. However, the hybridized approach based on low-level mise-en-scène features and high-level semantic attributes yield in the best perceived quality along all metrics ($p < 0.05$).

3.3.6 Discussion

Validity of Our Study

The internal validity of our study is supported by the accuracy of our research design and by the quality of study execution. We have carefully implemented various mechanisms to control the exactness of the tasks' execution. Obviously, the individuals' intrinsic characteristics and actual behavior always bring to an experiment a myriad of factors that can be hardly controlled [77]. In terms of external validity, the results of our study are limited to those participants and conditions used in our study. Moreover, most services accessible in the market provide a user experience fundamentally the same as the one utilized in our study, in terms of filtering criteria and information/navigation structures [77], and it is likely that replications of our study on other systems may lead to results consistent with our discoveries. At last, the high overall number of testers (100) allows us to generalize our outcomes to a wider population of users aged 20-50.

Research Questions

Our findings answer both of our research questions and show that mise-en-scène features combined with semantic attributes of movies improve both offline and online quality of recommendations. Recommender systems can leverage the gap between low-level and high-level movie features.

A finer grained examination of the statistically relevant relationships among all the diverse factors offers a significantly more articulated picture of the results, which demonstrate obviously contrasting results. More specifically, the two types of investigation (offline and online experiment) illustrate different pictures *w.r.t.* the impact of mise-en-scène features on novelty and diversity. Our explanation is that the low accuracy of visual-only recommendations negatively affects the user opinion on the other metrics. A possible interpretation of this result is to consider that previous studies confirmed a mismatch between offline and online quality of recommendations [77, 188].

3.3.7 Conclusions of study 2

This work represents a contribution to the research and study in the design of novel recommender systems, for the specific domain of movie recommendations and, from a more general perspective, video recommendations. Our research differs from previous work in this domain for a number of perspectives:

- We designed an online movie recommender system which integrates mise-en-scène features as a novel paradigm of movie recommendation, to be used for evaluation of recommendations with real users. In contrast, previous works on mise-en-scène features are based on only offline experiment.
- We compare three different CB recommendation approaches based on mise-en-scène features, semantic attributes and combination of both. Previous works limit their analysis to mise-en-scène features alone.
- Our results on the online evaluation of recommender systems based on mise-en-scène features (either alone or combined with semantic attributes) are totally new for the movie domain.

Overall, our findings extend our understanding of the potential of introducing mise-en-scène feature in movie recommendations to improve novelty and diversity of recommendations. The results of our study can be extended to other domains for recommending other multimedia products, such as music (*e.g.,* Spotify, and Pandora) and images (*e.g.,* Instagram, and Facebooks).

3.4 Study 3: "Hybridization with semantic-rich systems"

The results of this research work were published at following venues:

- As a long paper at the *Springer EC-WEB 2016* conference [91]
- As workshop paper at the *ACM Recsys 2016* conference [88]
- Finally, an extension of the work has been accepted at *International Journal of Multimedia Retrieval 2018* [92]

3.4.1 Introduction and related work

A major problem in RS is *the cold start* (CS) problem, *i.e.,* when a new user registers to the system or a new item is added to the catalog and no sufficient information is available about user and/or the item. In such a scenario, the system cannot properly recommend existing items to a new user or recommend a new item to the existing users, problems respectively known as *the new user* and *the new item* problems [17, 171, 280]. Another sub-problem of CS is the *sparsity* issue which happens when the number of available ratings is much lower than the number of possible ratings, which is particularly likely when the number of users and items is large. The effect of sparsity is that recommendations often become unreliable [171]. The inverse of the ratio between given and possible ratings is called sparsity. Typical values of sparsity are quite close to 100% in most real-world RS for example, the sparsity of the Yahoo! Music dataset and Netflix dataset of movies are 99.96% and 98.82% [106].

Different approaches have been proposed in the literature to tackle the CS problem, foremost *CB approaches*, *hybridization*, *cross-domain recommendation*, and *active learning* [58, 90, 113, 118]. Through this Phd dissertation, due to the relevance of the topic, we only address the only first two approaches, that is CB approach and hybridization.

CB techniques do not require the preference of other users for making recommendation. Therefore, as soon as some information about the user's own preferences/interaction with some items in the catalog are available, such techniques can be utilized to make recommendation especially in CS situations to cope with the sparsity issue. However, CBRS still suffer from the new-item problem. This is particularly the case when traditional metadata are used to serve as the content descriptors (features) since metadata production requires human effort. For example, imagine a video that has been just added to the YouTube catalog; how can a RS possibly recommend the video to users without having any information about the video or its content? In such scenarios, the proposed CB-approach in Study 1 that recommends items to users by extracting stylistic visual features from the video in an automatic manner (without human effort for annotation) and build a user-profile on implicit opinion of user on the stylistic aspects of videos can be used. The advantage of such an approach is that they can extract the descriptive attributes of items automatically from the item in order to serve as the item profile in a CBRS [89, 90, 93, 94]. As the result, in the most severe cases, when a new item is added to the catalog, CB methods can still generate recommendations, because they can extract multimedia features directly from the item and use them to make recommendations. It is noteworthy that while CF systems have CS problems both for new users and new items, CBRS (specially the ones using multimedia content analysis) have only CS problems for new users [19]. In this regard, while in Study 1 we focused our attention mainly on CB recommendation that use stylistic visual features to effect recommendation and showed that can be used to replace traditional metadata, in this chapter we would like to focus our attention on the second approach to combat CS problem that is hybridization.

An alternative technique to tackle the new item problem and maintain a higher level of recommendation accuracy is *hybridization.* A review of different hybrid and ensemble recommender systems can be found in [21,55]. Our main goal in this study is to study effective hybridization techniques in order to combine various CB and CF techniques so as to solve the abovementioned problems. We should mention that in Study 2, we also presented an study involving a hybrid RS. However, in that study our main goal was to compare and assess the result of an offline recommendation experiment with that of an online study under rather a simplistic hybridization approach which was built by interleaving the recommendation outputs of two CB systems. This is while in this study, we wish to have a more algorithmic approach toward the hybridization technique and study different SoA novel algorithms that can leverage the information originated from different RS. We present this section by studying hybrid system that combine two or several CB systems (Hybrid CB systems) and the ones combining a CB and CF system (Hybrid CB+ CF). We study two novel algorithms in each category (1) *canonical correlation analysis (CCA)* for combining two CBRS on the feature-side and (2) *factorization machines* (FM) to build a hybrid CB+CF.

3.4.2 Hybridization using Canonical Correlation Analysis

Multimedia features can be classified according to various dimensions. One of the most important dimensions from a human perspective is their *semantic expressiveness* (see section 2.1). It is common to distinguish three levels, with increasing extent of semantic meaning:

- *Low-level:* Low-level features are close to the raw signal *e.g.,* energy of an audio signal, colors in an image, motion trajectory in a video, or number of characters in a text.
- *Mid-level:* Mid-level are often expressed as a combination or transformation between different low-level features. For example they can be obtained by applying human auditory models to the amplitude or

frequency representation of an audio signal, or they are inferred from low-level features via machine learning. These features are more advanced than low-level ones, but farther away from being semantically meaningful than high-level ones.

- *High-level:* High-level features are close to human perception and interpretation of the signal (*e.g.,* motif in a classical music piece, emotions evoked by a photograph, meaning of a particular video scene, story told by a book author)

Classical approaches to multimedia recommendation use single modality among different modalities (audio, visual, textual). Users use textual queries in order to search text repositories, visual queries to search for image databases and so forth. Unimodal modeling keeps the semantic exploration of the content (from a semantic expressiveness point of view) limited. In addition, in the modern information era of today, this paradigm is of less used since multimedia content is ubiquitous. As the result, in the modern literature of multimedia, multimodal modeling, representation, and retrieval have been largely addressed [247, 258].

To this end, in this study we investigate a hypothesis which is built around the notion that explicit modeling of correlations between images and text is important and that by accounting for cross-modal correlations, we can enhance the informativeness of the final descriptor toward improving the performance of the joint model compared to cases where the two modalities are modeled independently. In particular, we propose a *multimodal* fusion paradigm which builds a content model by exploiting the low-level correlation between visual and textual modalities. The method is named *canonical correlation analysis* (CCA) which belongs to a wider family of multimodal subspace learning methods known as *correlation matching* [154, 247]. The approach is different with classical multimodal video recommender systems such as [229, 339] in which the authors treat the fusion problem as a basic linear modeling problem without investigating the underlying feature spaces. Instead, the proposed fusion method based on CCA is a technique for joint dimensionality reduction across two (or more) feature spaces in such a way that it can provide heterogeneous representations of the same data and maximize the correlation between the two. Although there has been extensive studies on exploiting correlation matching in the multimedia community (retrieval, representation), we do not know of a work to date in the RS community which leverages such methods in the multimedia domain. We articulate the research question of the this study as following:

Combining the stylistic visual features (low-level) with metadata (high-level) features by a fusion method which leverages the underlying feature space and maximizes the pairwise correlation can lead to more accurate recommendations, in comparison to recommendations based on the features when used in isolation.

The main contributions of this study are as following:

- We propose a novel fusion method to combine two semantically different set of features, one based on stylistic visual features and textual metadata. Unlike traditional fusion methods which do not exploit the relationship between two set of features coming from two different heterogeneous sources, the proposed method based on CCA maximizes the pairwise correlation between two different sets in order to better leverage the informativeness of the joint set. The results is improved recommendation quality compared to each of modalities when used in isolation.

- We evaluate our proposed technique with a large dataset with more than 13K movies that has been extensively analyzed (see previous studies) in order to extract the stylistic visual features.

Canonical correlation analysis

To enhance the performance of two different types of CBRS, the one based stylistic visual features (see Study 1) and the one based on textual metadata (tag), we propose a hybridization approach based on CCA that is extensively studied in the field of multimedia retrieval. The method is a popular method in multi-data processing and is mainly used to analyze the relationships between two sets of heterogeneous feature sets [136, 141, 247].

Let us assume $X \in R^{p \times n}$ and $Y \in R^{q \times n}$ be two set of features in which p and q are the dimension of features extracted from the n items. Let $S_{xx} \in R^{p \times p}$ and $S_{yy} \in R^{q \times q}$ be the *between-set* and $S_{xy} \in R^{p \times q}$ be the *within-set* covariance matrix. Also let us define $S \in R^{(p+q) \times (p+q)}$ as the *overall covariance matrix* - a complete matrix which contains information about association between pairs of features- represented as following

$$S = \begin{pmatrix} \operatorname{cov}(x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{3.8}$$

The aim of CCA is to identify a linear transformation represented by $X^* = W_x^T.X$ and $Y^* = W_y^T.Y$ so that it maximizes the pair-wise correlation across two feature set as given by Eq. 3.9

$$\arg\max_{W_x,W_y} \operatorname{corr}(X^*, Y^*) = \frac{\operatorname{cov}(X^*, Y^*)}{\operatorname{var}(X^*).\operatorname{var}(Y^*)} \tag{3.9}$$

where $\operatorname{cov}(X^*, Y^*) = W_x^T S_{xy} W_y$ and $\operatorname{var}(X^*) = W_x^T S_{xx} W_x$ and $\operatorname{var}(Y^*) = W_y^T S_{yy} W_y$. In order to solve the above optimization problem, we adopt the maximization procedure described in [136, 141] and solve the eigenvalue equation

$$\begin{cases} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \hat{W}_x = \Lambda^2 \hat{W}_x \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \hat{W}_y = \Lambda^2 \hat{W}_y \end{cases} \tag{3.10}$$

where $W_x, W_y \in R^{p\times d}$ are the eigenvectors and Λ^2 is the diagonal matrix of eigenvalues or squares of the *canonical correlations* and $d = rank(S_{xy}) \leq \min(n, p, q)$ is the number of non-zero eigenvalues in each equation. By computing $X^*, Y^* \in R^{d\times n}$, the fusion can be performed in two manners: (1) concatenation (2) summation of the transformed features defined by W_x and W_y:

$$Z^{ccat} = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^T.X \\ W_y^T.Y \end{pmatrix} \tag{3.11}$$

and

$$Z^{sum} = X^* + Y^* = W_x^T.X + W_y^T.Y \tag{3.12}$$

3.4.3 Hybridization using Factorization Machines

In a second study about hybridization, we show how to use stylistic visual features extracted automatically from video files as side information to a CF system resulting to improve the overall quality of recommendation. We propose a hybrid system based on CF system based on Factorization Machines (FM). FM can mimic the most successful approaches in recommender systems including matrix factorization, SVD++ or PITF [265]. We show FM can be applied to include stylistic visual information as side information and to improve the overall quality of recommendation in comparison with pure CB or CF system alone.

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 mise-en-scène 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

Factorization machine

Factorization machines (FM) [261] can be seen as a generalization of support vector machines (SVM) and factorization models (*e.g.,* matrix factorization (MF) or tensor factorization) with the merit of combining the advantages of both systems. The key advantage of FM compared with SVM and MF include: (1) FM can serve as a general predictor capable of working with real-valued feature vectors (as for SVM) under huge sparsity (not working for SVM); (2) FM learn the latent factors for all variables (not only user item interaction as in MF), including side features, thus allowing for interactions between all pairs of variables. For this reason FM are known to be capable of modeling complex relationships in the data. As stated in [297], the key distinctive idea of FM is to transform user-item interaction (preferences) and the associated side information into a real-valued feature vector and train an SVM-like learning algorithm for regression against the unknown preferences. We formally presented the model of FM in Section 2.1.

3.4.4 Experimental Results for study A and B

Study A

In this sub-study of the hybridization section, we present the experimental details and results of the evaluation. For this purpose, we used the CBRS based on "k-nearest neighbor" defined by

$$\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{j \in N_u(i)} s_{ij}} \sum_{j \in N_u(i)} s_{ij}\, \mathbf{r}_{ui} \tag{3.13}$$

where $\hat{\mathbf{r}}_{ui}$ is the unknown preference score (*i.e.,* rating) for user u and item i, $N_u(i)$ is the items rated by user u most similar to item i and s_{ij} is the similarity score between items i and j (the content-centric similarity). We fed the CBRS with three types of content features: (1) stylistic visual features presented in Study 1 [90]; (2) tag features represented by latent semantic analysis (LSA); (3) the proposed visual-textual descriptor using CCA. We used the movielens dataset whose characteristic resembles that in Study 2.

Table 3.9: *Quality of recommendation w.r.t Recall, Precision and MAP when using low-level visual features and high-level metadata features in isolation compared with fused features using our proposed method based on Canonical Correlation Analysis.*

Features	**Fusion**	**Recall**		**Precision**		**MAP**	
	Method	@5	@10	@5	@10	@5	@10
textual (tag)	-	0.0028	0.0049	0.0045	0.0041	0.0025	0.0021
visual (stylistic)	-	0.0038	0.0046	0.0051	0.0037	0.0035	0.0028
visual + textual	CCA-sum	0.0055	0.0085	0.0081	0.0069	0.0045	0.0038
visual + textual	CCA-ccat	**0.0115**	**0.0166**	**0.0140**	**0.0115**	**0.0091**	**0.0080**

The results of the experiments are presented in Table 3.9. As it can be noted, both of the hybrid methods (*i.e.,* the one based on CCA-sum and CCA-ccat) outperform the textual and visual CB alone *w.r.t.* all evluation metrics presented and across different cutoff values, *i.e.,* @5 and @10. For example, in terms of Recall@5 and @10, it can be seen that hybrid CCA-ccat obtains a score of 0.0115 and 0.0166 respectively which is significantly higher than textual with scores 0.0028 and 0.0049 and/or visual with values 0.0038 and 0.0046, when used in isolation. It can be also seen that even the second variation of the hybrid method based on CCA-sum can reach scores of 0.0055 and 0.0085 for Recall@5 and @10 which is still higher than individual textual or visual content features, even if not significantly. These scores indicate that by combining two heterogeneous textual and visual feature sets using CCA we can leverage the informativeness of each set and obtain a more informative joint visual-textual descriptor which is considerably better than recommendation based on these content features individually when considering recall.

In terms of precision, the same trend of results can be noticed. For example, the hybrid approaches obtained scores of 0.0140, 0.0115 for CCA-ccat *w.r.t.* Precision@5 and Precision@10. These values are significantly higher than individual CB methods with scores of 0.0051, 0.0037 for visual and 0.0045, 0.0041 for textual. The alternative fusion method (CCA-sum) obtained precision scores of 0.0081, 0.0069 which is still better than the other two individual baselines.

The results are consistent *w.r.t.* MAP hence fusion method based on CCA-ccat producing the best quality, obtaining scores of 0.0091, 0.0080 for MAP@5 and MAP@10 significantly better than the textual with scores 0.0025 and 0.0021 and visual with scores 0.0035 and 0.0028 respectively. Accordingly, the fusion of the stylistic visual features and textual tag provide the best performance in terms of the MAP metric.

Overall, the results validates our hypothesis that by combining the visual features extracted from movies with tag content and using the fusion method based on canonical correlation analysis which maximizes the pairwise correlation between heterogeneous features set, we can significantly improve the overall quality of recommendations. This is promising outcome and shows the great potential of exploiting stylistic visual features together with other sources of content information such as tags in generation of relevant personalized recommendation in multimedia domain.

Study B

In order to generate recommendations using our low-level visual features, we adopted the FM [261] method as described in the previous section. We fed the model with two types of content features serving as item side-information. The content features include the one based on stylistic visual features (introduced in Study 1) and genre features (containing 19 genre labels). We also added top-rated recommender to serve as a non-personalized recommender baseline. In this line, FM can compute rating predictions as a weighted combination latent factors, low-level visual features/genre labels, biases, thereby capable of capturing complicated relationships in the data.

We used the user-rating matrix dataset based on the Movielens dataset whose characteristics match the one in previous study. As for the genre labels, the movies are classified into 19 genre labels: *Action*, *Adventure*, *Animation*, *Children's Comedy*, *Crime*, *Documentary*, *Drama*, *Fantasy*, *Film-Noir*, *Horror*, *Musical*, *Mystery*, *Romance*, *Sci-Fi*, *Thriller*, *War*, *Western*, and *unknown*.

The first experiment entails applying a number of normalization techniques that would help improvement of the modeling using visual features. Our idea was that since the visual features are real-valued features and given that FM are designed to work also with real-valued features, different normalization of the data can result in different performance of these features. We tried different normalization methods in order to find out which one fits best to our data as described in the following

- *Log-Max norm*: for every feature value, it is scaled by passing though a logarithmic scaling function (natural logarithm). This would change the distributions to be approximately normal. We then performed max normalization in order to fit the values of each feature within the range of 0-1.

- *Quantile-Max norm*: for every feature value, the values are normalized by applying quantile normalization [46]. This would transform the distribution of all the features to be similar to each other. The procedure is followed by a max normalization step in order to scale the values of each feature to be within the range of 0-1.

- *Log-Quantile-Max norm*: for every feature value, the values are scaled by passing them though a logarithmic transformation followed by a quantile normalization. Again, the max normalization is performed in order to scale the values to fit in the range 0-1.

Details for the normalization results can be found in [88]. The best results were obtained by using Log-Quantile-Max norm. Here we show only the final results under best normalization setting for visual features. Table 3.10 presents the final results of the evaluation. Similar to study 3-A, we tested the quality of recommendation *w.r.t.* three different metrics, Precision and Recall but chose F1 instead of MAP. From the results, it can be seen that in terms of Precision, Recall and F1 the the stylistic visual feature performs the best across different cut-off values @5 and @10.

For example *w.r.t.* precision, the best technique our proposed visual FM achives scores of 0.0367 and 0.0343 for Precision@5 and Precision@10 respectively. This is while genre-based FM obtained scores of 0.0041, 0.0038. This result is promising since it shows that adoption of FM using visual features as side-item information can result in precision scores much better than genre-based recommendation.

Similar trend of result has been observed *w.r.t.* recall metric. Here our proposed visual FM obtained scores of 0.0272, 0.0488 for Recall@5 and @10 well-above genre FM which is 0.0025, 0.0049.

We also computed the F1 metric which is a harmonic mean of precision and recall. nd combines both metrics in a single metric. A nice property of the harmonic mean is that it substantially increases or decreases if both of the components increase or decrease. In reality, it is difficult to improve F1 by maximizing either precision or recall, the reason we chose it in this study.

Comparing the results, our proposed visual FM outperforms all the other technique in terms of F1 as well. It achieved the F1 scores of 0.0312, 0.0403 which is significantly higher than genreFM baseline with scores 0.0031, 0.0043 for F1@5 and F1@10.

As expected, in all the above cases, the non-personalized recommendation baseline based on top-rated recommendation technique is the worst technique among all in terms of precision recall and F1.

Comparing all these 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.

These are promising results indicating that Factorization Machines fed with the low-level visual features achieves much superior predictive power than when it is fed with the genre, in predicting the relevance of movies for users.

Table 3.10: *Performance comparison of various recommendation techniques*

Algorithm	**Precision**		**Recall**		**F1**	
	@5	@10	@5	@10	@5	@10
Visual FM	**0.0367**	**0.0343**	**0.0272**	**0.0488**	**0.0312**	**0.0403**
Genre FM	0.0041	0.0038	0.0025	0.0049	0.0031	0.0043
Top rated	1.390e-05	1.042e-05	1.922e-06	3.087e-06	3.377e-06	4.764e-06

3.4.5 Conclusion of study 3

This study presents a novel approach in the domain of content-based movie recommendations. It particularly addresses the under-researched domain of movie recommendation by combining visual properties obtained from videos with semantic rich sources of information such as metadata and collaborative knowledge shared in CF system. Both of the latter sources of information are semantically rich sources since they are result of human knowledge as opposed to visual features which are extracted in a completely unsupervised manner. This study investigates adoption of two state of the art hybridization techniques based on Canonical Correlation Analysis (CCA) and Factorization Machines (FM) for leveraging the joint information obtained from visual and semantic-rich sources. The results in both studies showed significant improvement of recommendation utility along different accuracy metrics.

3.5 Conclusion of Chapter 3

CHAPTER 4

Multimodal Content-Based Video Recommendation

4.1 Introduction

Video recordings are intricate audio-visual signals. When we watch videos, we can effortlessly record considerable details communicated to us through different multimedia channels, in particular, the audio and visual channels. Accordingly, the video content can be described in versatile manners since its perception is not restricted to one view. Such multiple facets can be unveiled by descriptors of visual and audio content, but also in terms of metadata, either editorial, *e.g.,* genre, cast, director, or by leveraging the wisdom of the crowd, *e.g.,* tags, reviews.

Content-based movie recommender systems (CBMRS) traditionally base their recommendations on metadata since they are usually assumed to describe the semantics of video content. On the other hand, CB multimedia descriptors automatically extracted from the audio-visual signals provide a complementary information to identify videos that "look similar" or "sound similar". These discerning characteristics of multimedia meet users' different information needs.

Movie recommendation systems (MRS) are conventionally powered by either collaborative filtering (CF) or content-based filtering (CBF) algorithms [190, 214, 268]. While CF's assumption is that the target user (to whom the recommendation is effected) would prefer content similar to the content other like-minded users prefer, CB assumes that the target user would prefer items similar to those he or she liked in the past. More particularly, CF methods leverage the fact that the specified ratings are often highly correlated across users and items, therefore the unspecified ratings (to be predicted by the RS) can be imputed by considering the *inter-item* correlations (*item-based* CF model) or *inter-user* correlations (*user-based* CF model). Some models use both types of correlations. CB models combine the preference indications of a single target user and content information available about the items in order to build a user model (aka *user profile*) and compare the user profile with the descriptive information of the content (aka *item profile*) in order to effect recommendations. In contrast to CF, CB recommenders solely require target user's own preference in order to make recommendation making them suitable in cold-start situations [19, 214].

However, the extent to which CB is used, and in turn the sophistication of respective algorithms, strongly varies between different domains, *i.e.,* the type of items recommended. While in the multimedia community, extracting descriptive multimedia features from different media content such as text, audio, image, and video content is a well-established research task, the RS community — for a long time driven by movie recommendation tasks — even has a different interpretation of the term "content". CBF approaches frequently resort to (high-level) metadata only, *e.g.,* title, tags, actors, or plot of a movie, as the single source for content-based recommendation models, thereby disregarding the wealth of information encoded in the actual audio visual signals.

Addressing this research gap, in this work we address specifically this issue with the help of machine-based processing of multi-modal information. In particular, the main contributions of this endeavor is the proposal of a *multi-modal content-based recommender system* (CBRS) that adopts latest state-of-the-art visual and audio features along with metadata, in order to build *rich item descriptions*. We refer to this rich content information as the *Video Genome*, since it can be considered as the footprint of both content and style of a video (similar to a biological DNA omposed of long sequences of four letters A, T, C, G and referred to as nucleotides) [52].

There are several other motivations for the use of audio-visual features in CBRS as opposed to metadata. Metadata are prone to errors (in particular being user-generated) and labor-intensive/expensive to collect (especially the expert-generated ones). User-generated metadata often exhibit user or community biases and might therefore not fully, or only in a distorted way, reflect the characteristics of a video [61, 197]. Furthermore, metadata are often rare or absent for new videos, making it practically infeasible to generate high-quality recommendations in CS scenarios [1].

Even in scenarios where metadata are available in abundance, due to their unstructured or semi-structured nature, often complex natural language processing (NLP) tasks are required for pre-processing, *e.g.,* syntactic and semantic analysis, stemming, or topic modeling [19]. One last motivation for the use of multimedia features in CBRS is that the perception of a movie in the viewer's mind is influenced by many factors, not only related to the genre, cast, and plot, but also to the overall film style [47]. Although may not be consciously noticed, these factors play a key role in driving the viewers' experience. For example, two movies may be from the same genre and director, but they can be different based on the movie style. To give an example, "Empire of the Sun" and "Schindler's List" are both dramatic movies directed by Steven Spielberg and depicting historical events. However, they have a completely different film style with the former being shot like a documentary in black and white and the latter using bright colors and making excessive use of special effects. Although very much similar *w.r.t.* traditional metadata (director, genre, year of production), these two movies are different based on their styles and are likely to influence the viewers' feelings and emotions differently [90].

To this end, in this chapter, we aim to answer the following research questions:

RQ1: *Can the exploitation of the* video genome *representing rich item description provide a superior recommendations quality compared with traditional approaches that use human-generated metadata?*

RQ2: *Which of the video multimedia channels (visual or audio) play a more significant role in driving users' preference toward a video?*

RQ3: *To which degree hybridizing audio and visual information can improve the quality of recommendation?*

In answering these research questions, we particularly address the abovementioned imperfections of exclusively metadata-based MRS, exploiting multi-modal signals in various ways.

The results of this research study is under review at the Journal of User

[1] In RS, cold-start refers to the situation where rating and/or metadata are rare or absent. Here we use the terminology to refer to the lack of metadata as CBRS do not require the ratings of other users to generate recommendations [269].

Modeling and User-Adapted Interaction (UMUAI) at the time this thesis was submitted. In addition, preliminary results of the work can be found in [84].

4.2 Related Work

Since the present research work, bridges the research fields of *multimedia* and *recommender systems*, we organize related work according to the perspectives and advances in these two research fields. In what follows, we primary present the state-of-the-art in multimedia recommender systems, subsequently investigating the background of the multimedia descriptors used in this chapter and their successful applications in other disciplines.

4.2.1 Multimedia Recommender Systems

Generally speaking, the term multimedia can refer to the *media content* that can be offered as a recommendation to users or as a modality from which various features (audio, visual, textual) are extracted.

As for the former view, a multimedia object (aka multimedia item or multimedia document) can be seen as composition of one or several media contents for examples a video clip, a book, a Website or a music piece. In addressing this view of multimedia, we only consider research works implementing a CB or hybrid system containing two or several CB systems and/or CB + CF. We intentionally refrain from CF models since our motivation is to improve our understanding on the contribution of the individual and combined multimedia features, *w.r.t.* various recommendation evaluation metrics (*e.g.,* accuracy and beyond accuracy metrics). Given that pure CF system do not require CB descriptors to generate recommendations and that they utilize community preferences/ratings (see Chapter 2) as their main source of information, we initially refrain from discussing CF-based multimedia recommender systems. In this regard, it is of our belief that to date, majority of attention in the community of RS has been paid on music and image domain and to a much lesser extent to video recommendation using mulimodal content analysis/features. For instance, while acoustic features play a key role in music recommendation research for years, multimedia content analysis is highly *under-researched* in the video domain. Nevertheless, in the following, we review some of the most closest relevant research works to the present work and shed their differences with the current study.

As one of the first research works [336] proposed a multimedia RS (image, video and textual document) using eye-tracking technologies as a novel paradigm for inferring user's taste over multimedia documents. By extracting descriptive features from the multimedia documents and interpreting user's gaze attention (time and location), the authors proposed a new paradigm for user preference inference and transferring this into a user profile model in order to improve the quality of CB recommendation. However, the multimedia features used in this research work are limited to the visual modality (unimodal) and deprecated, *i.e.,* "Auto Color Correlogram" [156].

[229, 230, 339] proposed a video RS named "Video Reach" which receives as input an online video and related information (query, title, tags, and surrounding text) and recommend as output videos which are relevant in terms of multimodal relevance and user feedback. Two types of users' feedback are exploited: (i) Browsing behavior and (ii) Playback on different potions of the video (only specific to [229]). Although these research works (by the same authors) provide interesting insights for using fusion schemes to build a hybrid (multi-modal) RS and basing this on user behavior, they have several shortcomings as well.

Firstly, the proposed fusion scheme in "Video Reach" is based on the Attention Fusion Function which filters out videos based on their textual similarities and visual similarity in a second stage. Adoption of such a sequential approach to find similar videos may not correspond to users' differing criteria to assess a video as relevant/interesting and may risk loosing important information. Secondly, the visual feature employed in this research work is limited to color histogram which is quite basic and naiive. Today, there exists a much wider and richer set of visual features that can be used for the recommendation process. Thirdly, it is not clear how the visual feature are temporally aggregated to build a video-level descriptor. Fourthly, an empirical set of weights are selected to define intra-modality and inter-modality coefficients in linear model. The weights associated with textual term are given a higher value making the assumption a-priori that textual keywords are more relevant compared with the visual and aural keywords, without investigating the opposite. Although the authors show that this assumption is adequate to initial the recommendation and then adjusting weights incrementally, it is not clear nor elaborated what the effect of such an empirical assumption is.

[36] proposed a multimedia (image — video — document) recommender platform designed for the Cultural Heritage domain. The proposed RS is context-aware recommender system (CARS) which combines uniformly heterogeneous multimedia data as content information. The focous of this research work is on contextualization although it used multimedia

content in the recommendation process. One of the key assumption stated in this work is that high-level metadata and low-level features (extracted in an automatic or semi-automatic manner) are correlated. However, as reported by the authors themselves, for some particular content, this correlation assumption may not hold true.

As a general conclusion, from the existing few video recommendation approaches, most of the works are powered solely by metadata or shyly integrate basic content descriptors from single modality and combination of those.

4.2.2 Multimedia Content Description

Multimedia features have been largely exploited in tasks other than movie recommendation.

Audio. In the audio and music domain, CB descriptors play a crucial role for a wide rang of tasks from music piece/song recommendation [44, 173, 259] to music audio similarity estimation [38, 45, 290], genre classification [203, 220, 309], emotion recognition [283, 305, 340], and/or semantic tagging (auto-tagging) [41, 218, 307]. The type of audio CB features used in these tasks include but not limited to: *timbral features* (describing the spectral shape of the audio signal), *tonal features* (describing pitch classes, key, chords, *etc.,*), and *temporal descriptors* (describing beats, onset rate, rhythm and forth). The Mel Frequency Cepstral Coefficients (MFCC) [212] from the timbral feature category is the most extensively adopted descriptor in many audio and music tasks.

Recently, *i-vectors* have gained increasing attention since they are capable of capturing variety of information from the speech signal such as the language [222], and speaker-related characteristics such as accent [33], emotions [333], the age [32] and the identity [82] of the speaker. Despite this, the applications of i-vectors are not limited to speech and they have been used as the state-of-the-art representation in various in audio-based music information retrieval tasks, for instance: music similarity estimation [108], music artist classification [109], and singing language identification [195]. Recently i-vector features have also shown promising effect in playlist continuation to deal with the cold-start problem [316]. As yet another related field beside music, i-vectors are state-of-the-art features in the field of acoustic scene analysis [323] for the task of acoustic scene classification [107].

The second class of audio-based features we investigate in this study are named BLFs which also showed encouraging results for different music-related tasks such as music similarity estimation [290], genre classification [287,291], auto-tagging [286], and predicting the origin of music pieces [279].

Visual. In the visual domain, content-based descriptors can be computed/extracted based on different attributes of an image level for instance *color*, *texture* and *edges*. Such attributes of the image are widely used for tasks which involve processing of images or video data, including object, human and face detection [79, 322] and recognition [26], shot segmentation [30] and content based image retrieval [350] to name a few. Perhaps the most important breakthrough in this domain and in recent years has been the introduction of deep neural networks, that gained state-of-the-art status in an increasing number of applications, and in some of them even surpassed human expertise. Examples of such applications include but not limited to optical character recognition [51], lip reading [74] and image classification [144, 194]. With the advent of computer vision techniques, these CB descriptors and respective deep neural network algorithms have been used in subjective tasks as well. The main difference in these tasks in that the ground truth data is more human oriented and content can differ based on user preference. Some of these subjective task are media aesthetics [174], violent scene detection [124, 159], affective content prediction [246] and media virality [28].

In the light of the superior performances in the ILSVRC 2012 object recognition competition[2], the AlexNet [194] network was adopted in many different tasks either through re-training the network, modifying it or using directly the network for some other specific purposes and/or through recording the outputs of some of its layers as features. For example, it has been successfully adopted in tasks such as aesthetic ranking [186], emotion recognition in videos [204, 334], prediction of multimedia interestingness [98, 294], and text-to-video translation [319]. Specifically, we can report a set of aesthetic visual features gathered by [135], based on the works of [80, 174, 202] which have been used effectively in highly subjective tasks like prediction of media interestingness [75] and image popularity [103]. For the matter of the aesthetic aspects involved in thse descriptors, they can be seen as valuable assets for a RS which is consequently investigated in this study.

Multimodal. In order to combine various information sources effectively, one has to employ fusion strategies. There are two main fusion approaches: *early fusion* in which the system combines the descriptors at the input of the system and uses the joint-descriptor as the input to the system, *late fusion* [301] which combines either intermediate outputs of several sys-

[2] (http://www.image-net.org/challenges/LSVRC/)

tems run on different descriptors or their final outputs. Such fusion strategies are based on the hypothesis that a joint decision from multiple sources can be superior to a decision from a single source. In this line, while late fusion focuses on the individual strength of modalities, early fusion uses the correlation of features in the mixed feature space. There is no clear supremacy of one technique over the other as both show state-of-the-art (SoA) results on different scenarios.

4.2.3 Contributions

In this study, we propose a novel SoA multimedia recommender system which pushes forward the advances of the field in the following directions:

1. We propose a multi-modal CBRS which uses established/traditional multimedia features namely: *aesthetic visual features and audio block-level features* and novel SoA features based on *deep visual features and i-vectors audio features*. An advantage of the proposed system beside using content descriptors automatically extractable is that it uses as input movie trailers instead of the entire movies, which makes it more versatile and effective since trailers are more readily accessible than full-length movies. We show that the proposed system outperforms the traditional CBRS based on metadata (genre and tags). To the best of our knowledge, this is a novel approach since current CBRS are limited either to usage of single modality [87,88,90] or deprecated descriptors [229,339];

2. We propose a practical solution to the cold-start problem, where metadata are not available or are only partially available, with promising results;

3. We propose a rank aggregation strategy extending the Borda count approach [34] to fuse recommendations from different (heterogeneous) sources into a unified ranking of recommended movies, outperforming the results obtained in the conventional approach using metadata;

4. We present a comprehensive evaluation of the proposed CBRS framework via two wide and articulated empirical studies: (i) *offline experiment with simulated users*: a system-centric experiment to measure the simulated quality of recommendations in terms of accuracy-related evaluation metrics such as precision, recall, mean average precision and mean reciprocal rank as well as beyond-accuracy metrics such as novelty, diversity, and coverage [170]; (ii) *online study involving real*

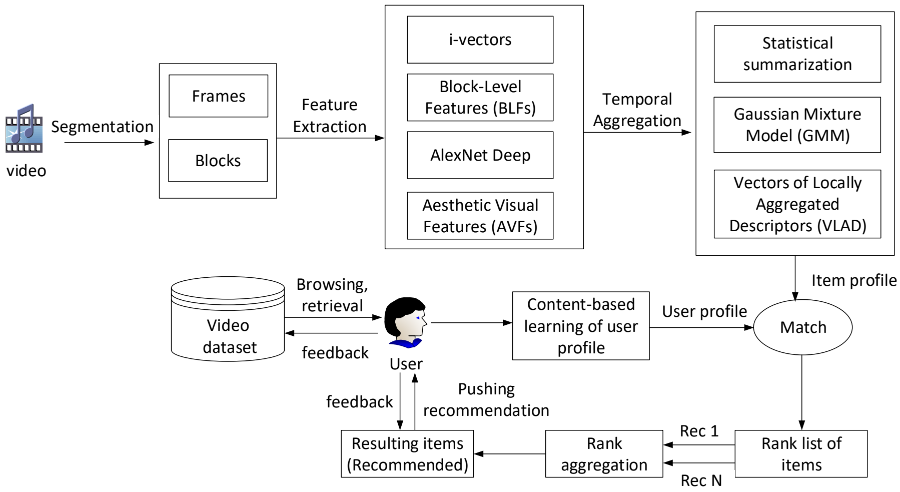


Figure 4.1: *General framework of a content-based recommender system for movies using multimedia content analysis.*

users: user-centric experiment involving 82 users, measuring the perceived quality of recommendation *w.r.t.* different subjective metrics such as perceived relevance, satisfaction, diversity and novelty.

4.3 Proposed Recommendation Scheme

The proposed content-based movie recommender system (CBMRS) includes several stages as illustrated in Figure 4.1. As mentioned earlier, the system takes as input movie trailers which makes it more flexible since trailers are more readily available compared to full-length movies. The steps adopted are described in the following:

The original size of the video dataset is hundreds of gigabytes, and covers over 20 days of video. Processing such amount of information is inefficient. Therefore, as the first step videos are segmented into smaller units. For the visual channel, motivated by the recent Google YouTube8M paper [14], the video frames are captured at rate of 1 fps. For the audio channel, we adopted *frame-level* and *block-level* segmentation as commonly applied in music information retrieval. The second step consists of extracting/computing meaningful content descriptors. Two categories of features are considered (see Section 4.4): (i) *multimedia*: composed of *audio* and *visual* features *metadata*: composed of movie *genres* and *user-generated keywords* (aka tags). Afterwards, temporally extracted features (*e.g.,* frame-level or block-level for audio) are aggregated to build a *video-level descriptor*. Different different video-level aggregation techniques are used such as statistical summarization, Gaussian mixture models (GMM) and vectors of locally aggregated descriptors (VLAD) [164]. Recommendations are generated using a CB system based on a standard k-nearest neighbor approach (see Section 4.5.2). In the final step, we propose an extended version of Borda count [34] method to fuse ranking results of different recommenders into a unified ranking of videos (see Section 4.5.3). The recommender are CB systems using different content descriptors.

These components are described in more detail in the following sections.

4.4 Rich Item Descriptions Defining the Video Genome

Similar to biological a DNA which is representative for a living individual, multimedia content information can be regarded as a genome *i.e.,* the footprint of content and style.

In this section, we present details about the content descriptors constituting the proposed *video genome*. The content descriptors are fundamental to the recommendation process and lie at the core of the video recommendation system. The proposed features were selected based on their successful applications in neighboring fields such as music and video information retrieval.

4.4.1 Audio Features

Many tasks in *multimedia information retrieval* depend on the extraction of low-level content features. The computational audio features considered here are inspired by the fields of speech processing and music information retrieval (MIR) and their successful application in MIR-related tasks, including music retrieval, music classification, and music recommendation [180]. We investigate two kind of audio features: (i) *I-vector features* which are frame-level representations computed from MFCCs acoustic features of audio segments. They provide a fixed-length and low-dimensional representation for audio excerpts containing rich acoustic information, and (ii) *block-level features* which consider larger segments of the audio signal known as *blocks*, typically a few seconds. Therefore, they can capture temporal aspects of an audio recording to a higher extent [180]. What both descriptors have in common is that they eventually model the feature at the level of the entire audio piece, by aggregating the individual feature vectors temporally (across time).

Block-level features: are extracted from large segments of audio recoding typically a few seconds. They have shown promising performance in audio and music retrieval and similarity tasks [286] and can be considered as state-of-the-art in this domain. The block-level feature framework [291] defines six features representing the following four aspects: (i) the spectral aspect (spectral pattern, delta spectral pattern, variance delta spectral pattern), (ii) the harmonic aspect (correlation pattern), (iii) the rhythmic aspects(logarithmic fluctuation pattern), (iv) and the tonal aspect (spectral contrast pattern). The feature extraction process in the block-level framework is illustrated in Figure 4.2. All block-level features use the same spectral representation based on the cent-scaled magnitude spectrum. For the input audio signal downsampled to 22KHz, it is transformed to the frequency domain by applying a Short Time Fourier Transform (STFT) using a window size of 2048 samples, a hop size of 512 samples and a Hanning window. The hop size helps to prevent from possible information loss due to windowing effect. Afterwards, the magnitude spectrum $\|X(f)\|$ with linear frequency resolution is computed and transformed into the logarithmic Cent scale. In a similar fashion, the magnitude spectrum $X(k)$ is mapped into a logarithmic scale. In order to make the computed spectrum invariant to variation of the intensity of the input audio, the mean of each sliding window is removed [288].

As shown in Figure 4.3, using the spectrogram of an audio recording, blocks with fixed length duration are selected and processed one at a time. The width of each block represents the number of temporally ordered feature vectors that a block contains. Features are computed within each block from which a global representation is created by aggregating the feature values computed across blocks using a summarization function, which is usually expressed as a percentile.

1. **Spectral pattern (SP):** The *spectral pattern* characterizes the frequency or timbral content of the piece of interest and is computed over the spectrograms of all blocks. Each block consists of 10 consecutive frames and a hop size of 5 frame is used, therefore there is half block overlapping between consecutive blocks. Frequencies are measured on the Cent scale, binned into 98 bands. Each frequency band is sorted within the block under investigation in order to obtain a time-invariant representation. For all blocks in the piece, the same procedure is repeated. This yields for each block a 98×10 matrix.

   Figure 4.4 compares the spectral pattern features for a classical piano piece by Shostakovich (Prelude and Fugue No. 2 in A Minor) and a

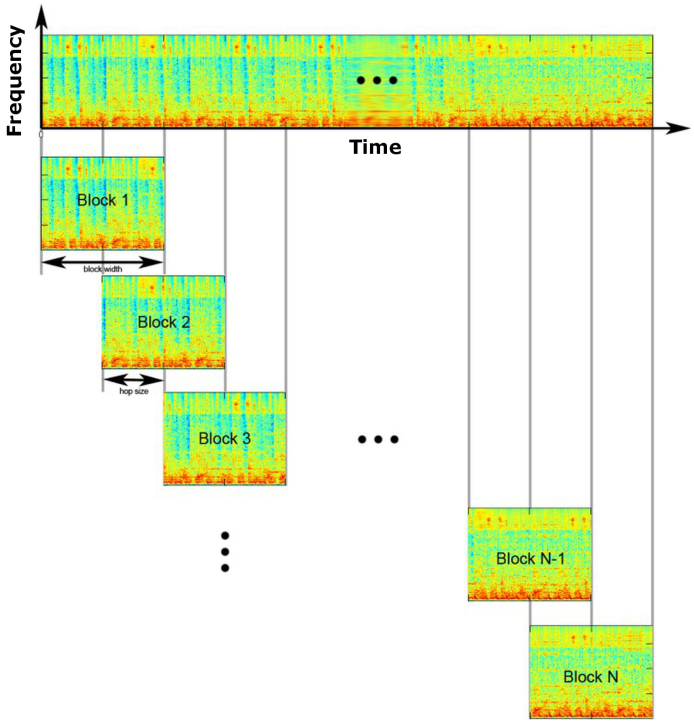


Figure 4.2: *Overview of the feature extraction process in the BLF.*

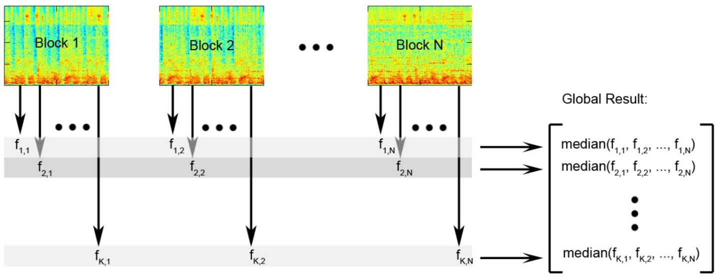
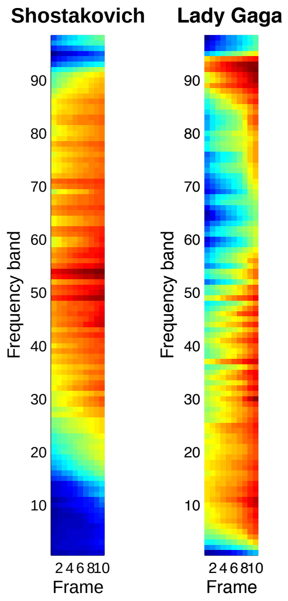

Block 1
Block 2
Block N
Global Result:
$f_{1,1}$
$f_{2,1}$
$f_{1,2}$
$f_{2,2}$
$f_{1,N}$
$f_{2,N}$
$f_{K,1}$
$f_{K,2}$
$f_{K,N}$
median($f_{1,1}$, $f_{1,2}$, ..., $f_{1,N}$)
median($f_{2,1}$, $f_{2,2}$, ..., $f_{2,N}$)
median($f_{K,1}$, $f_{K,2}$, ..., $f_{K,N}$)

Figure 4.3: *Obtaining a global feature representation from individual blocks in the block-level framework.*

pop song by Lady Gaga (Bad Romance). Note that within every frequency band, the energy values are sorted. As it can be seen, the two music pieces have quite different spectral characteristics in which the piano piece exhibits high activations in a limited number of frequency bands (45 to 55 Cents, i.e. $\approx$ 700 to 1600 Hertz) but is missing very



(a) *classical piece*(**b**) *pop piece*

Figure 4.4: *Spectral patterns for a classical piece (left) and a pop song (right).*

low frequencies. This is while, the pop song concentrates stronger activations in low bands (up to 25 Cents, i.e. ≈ 130 Hz) [180].

2. **Delta spectral pattern (DSP):** A few variants of the spectral pattern features are defined in the BLF in order to capture other characteristics of the spectrum. The *delta spectral pattern* captures note onsets [39] by detecting changes in the spectrogram over time. It perform this by measuring the difference between the original cent spectrum and a delayed version of the spectrum by 3 frames to emphasize onsets. In a similar fashion to SP, again each frequency band of a block is sorted. After this step, positive values in the difference spectra indicate note onsets, which are of interest. Ultimately, since the values within each band are sorted, the left-most columns of the 98×25 matrix are all negative and only the right-most 14 frames with largest values are taken into account where 14 frames is chosen to save some computational and memory resources [180]. Similar to SP, DSP uses 0.9-percentile as summarization function.

3. **Variance delta spectral pattern (VDSP):** The VDSP is identical to the DSP with one main difference that it uses the variance as summa-

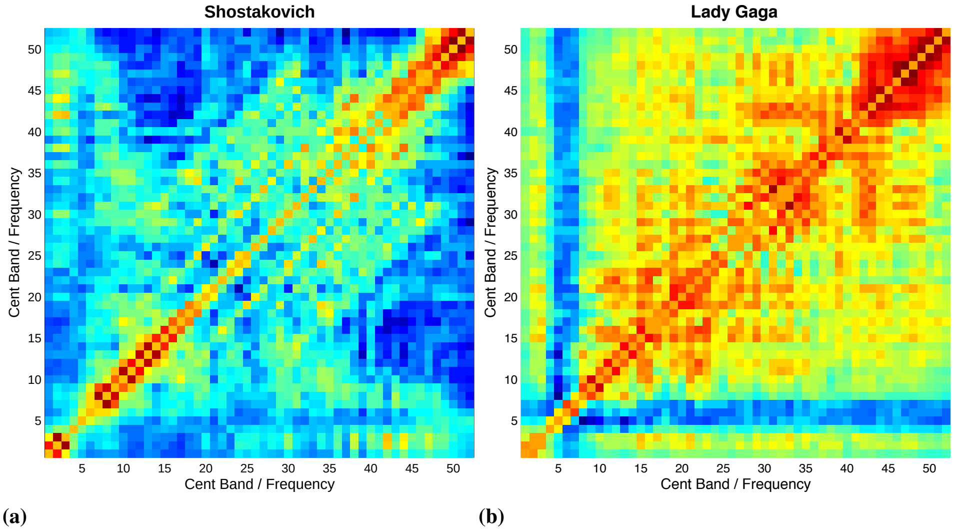


Figure 4.5: *Correlation patterns for a classical piece (left) and a pop song (right).*

rization function instead of the 0.9-percentile so as to capture variations in onset strength across time [180].

4. **Correlation pattern (CP)**: The CP captures the harmonic relations of frequency bands in the presence of sustained tones. In order to achieve this aim, the frequency resolution is first reduced to from 98 to 52 bands. Then, a pairwise linear correlation coefficient (Pearson Correlation) between each pair of frequency bands is calculated, giving rise to a symmetric correlation matrix. A block size of 256 frames with half-block overlapping between consecutive blocks is chosen. The result of computing correlation between all pairs of frequency bands is a 52×52 matrix for each block. A median summarization function is the used to aggregate all matrices within blocks. Figure 4.5 illustrates the difference between CP of a classical and a pop song. As it can be noted, the patterns of the classical song reveal harmonics (shown by orange and yellow squares) which remain outside of the diagonal. For the pop song however, there exit larger square areas of high correlations between adjacent frequency bands reflecting high interrelations between similar notes. As another example, if a bass drum is always hit simultaneously with a high-hat, this would yield a strong positive correlation between low and high frequency bands [180, 287].

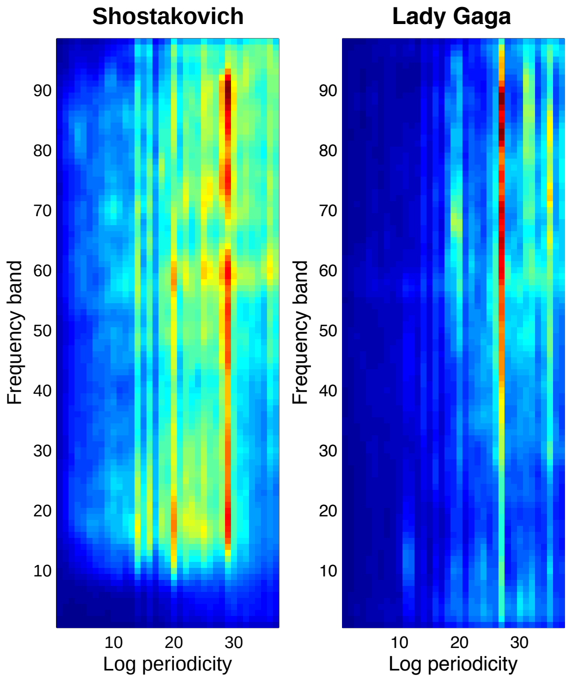
5. **Logarithmic fluctuation pattern (LFP):** The LFP is an extension

to Pampalk et al.'s fluctuation patterns feature [242] and represents the rhythmic structure of a song. First, the audio piece is segmented into a sequence of overlapping 6-second-segments and a fast Fourier transform (FFT) is computed for each segment to obtain the power spectrum to which several loudness transformations and psychoacoustic preprocessing techniques are applied. A second FFT is applied in order to captures information about the frequency structure. Applying the second FFT transforms time–frequency representation into a periodicity–frequency representation. The frequencies corresponding to the periodicities are frequency patterns repeated and repeated over time known as fluctuations. The fluctuations reveal the frequencies which reoccur at certain intervals within the 6-second-segment under investigations. The result of the process is computation of a 20×60 matrix that contains energy levels for 20 critical frequency bands over 60 bins of periodicities, ranging from 0 to 600 beats per minute. In a final step of psychoacoustic processing, order to account for different intensities the human ear perceives recurring beats at different periodicities, the amplitude modulation coefficients are weighted based on the psychoacoustic model as proposed in [243].

In contrast to the original fluctuation patterns [242], the LFP uses a logarithmic scale to model frequencies. Figure 4.6 illustrates the LFP for the piano piece and the pop song. As it can be observed, the pop piece shows a concentration of energy in very few periodicity bins while the classical piece exhibits several regions of high intensity.

6. **Spectral contrast pattern (SCP):** Finally, the SCP estimates the "toneness" of a spectral frame (the tonal aspect). This is realized by computing the differences between spectral peaks and valleys in several frequency sub-bands. Pronounced spectral peaks roughly correspond to tonal components, whereas flat spectral parts are often percussive or noise-like. In the BLF, the SCP is computed from a Cent scaled spectrum subdivided into 20 frequency bands. The difference between the minimum and maximum value of the frequency bins in each band, results in 20 spectral contrast values per frame. Similar to SP described above, the values belonging to an block are sorted within each frequency band. A block size of 40 frames and half-block (*i.e.,* 20 frames) overlapping between consecutive blocks is used. The summarization function is the 0.1 percentile.

I-vector features:



(a) *Classical piece* **(b)** *Pop song*

Figure 4.6: *Logarithmic fluctuation patterns for a classical piece (left) and a pop song (right).*

The i-vector approach got its fame first time in its successful application in the field of speaker verification with promising performance [82][3]. Ever since, it has been used as the SoA representation learning technique in different audio-related domains [83, 107, 316]

I-vector is a fixed-length and low-dimensional representation for audio excerpts containing rich acoustic information. It is usually extracted from short audio segments (of length 10 seconds to 5 minutes) of acoustic signals such as speech, music and acoustic scene. The standard approach to compute i-vector features are based on frame-level features such as Mel-Frequency Cepstral Coefficients (MFCCs). The procedure for extraction of i-vectors entails learning a Universal Background Model (UBM) to model the average distribution of all the audio data in the acoustic feature space. In order to learn such a universal model, a dataset containing sufficient amount of data of different audio is used. The learned UBM serves as a reference corpus to measure the amount of shift of each arriving audio segment from

[3] The task of verifying the claimed identity of a speaker based on the speech signal from the speaker [3]

the UBM. I-vector represents such a shift by a latent factor representation.

Formally speaking, let us define **total variability** for movies as the deviation of a video clip representation from the average representation of all the video clips. A UBM is learned using a Gaussian Mixture Model (GMM) on the acoustic features of segments. The mean vector of the learned model is high-dimensional representation for each segment. Afterwards, a Factor Analysis (FA) procedure is applied in order to capture the shift of the adapted model from the UBM. To apply the FA, the adapted GMM mean super-vector $\mathbf{M_s}$ – which is adapted to an audio segment s – is decomposed as follows

$$\mathbf{M_s} = m + \mathbf{T}.y_s \tag{4.1}$$

where m is the UBM mean super-vector and $\mathbf{T}.y_s$ is an offset that captures the shift from the UBM. In Eq.4.1,$\mathbf{T}$ is the factorization matrix called total variability matrix (TVS) which is usually learned via an expectation-maximization (EM) procedure [82, 224] and y_s is a low-dimensional latent variable known as *i-vector* learned via a maximum a posteriori (MAP) estimation procedure.

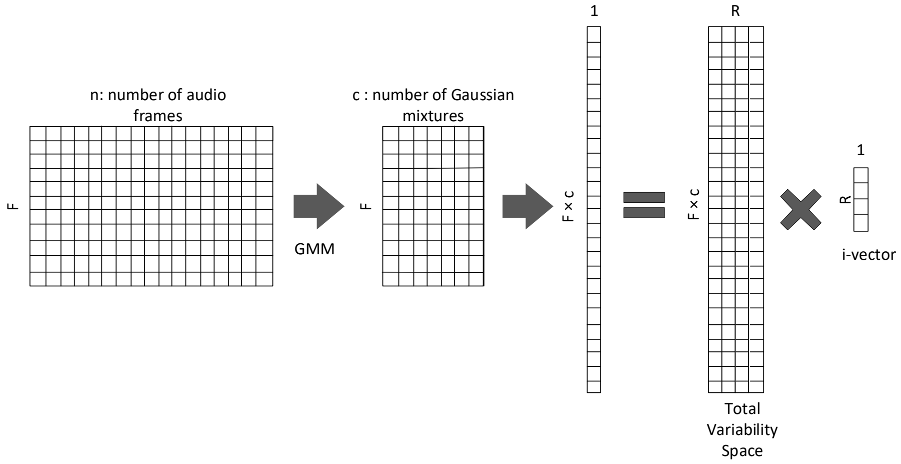


Figure 4.7: *Graphical representation of the procedure leading to i-vector factor analysis.*

A graphical illustration of different features extracted during i-vector FA is shown in Figure 4.7. Let us assume the frame-level features such as MFFCs are extracted where F is the dimension of the acoustic feature vectors. A GMM is trained to represents the distribution of feature vectors (the acoustic feature space) where C is the number of Gaussian components. Next, a GMM supervector is created by concatenating the mean vectors m_c

for each mixture component $c = 1, ..., C$. The dimensionality of the GMM supervector is therefore equal to $(F.C) \times 1$. Ultimately, the rectangular matrix T known as TVS is learned via EM procedure and used to extract i-vectors from the GMM (M_s in Eq. 4.1).

In Figure 4.8, we present a practical illustration of the i-vector extraction pipeline in a block-diagram representation. The pipeline start from frame-level feature extraction (*e.g.,* MFCCs), continues with i-vector extraction, an optional LDA step for supervised tasks and finally recommendation. It should be noted that unlike BLF in which features are extracted in the same manner from train and test, in i-vector the procedure used in train and test phases are slightly different. The framework consists of the following five stages:

1. *Frame-level feature extraction:* Typically, MFCCs are used in the feature exaction stage. They have been the dominant features in speech recognition. Their success owed to the ability of representing the spectrum envelope of speech signals in a compact form. Since the start of year 2000, they were also proven useful to model music and audio sounds [115, 213] and ever since they are the most dominant feature in the field of music information retrieval (MIR). Although, it is possible to use other features [310] for feature extraction, we stay ourselves aside from the feature engineering part and instead focus on the FA technique. In our experiment, we used 20-dimensional MFCC features;

2. *Computation of Baum-Welch statistics:* After extracting MFCCs, a set of statistics are computed for each audio known the the Baum-Welch (BW) statistics. The statistics are required to learn the bases for the total variability subspace in the subsequent i-vector extraction phase. In this step, a sequence of MFCC features are represented by the BW statistics (0-th and 1-st order Baum-Welch statistics) using a GMM as prior [176, 198].

3. *I-vector extraction:* In this stage, the subspace matrix T known as TVS is trained via EA. I-vectors are then extracted based on the estimated TVS. Both TVS and i-vector require statistics extracted from the training set.

   The extraction of i-vectors reduces the dimensionality of the movie clip representations and improves the representation for a recommendation task;

4. *LDA Supervised task:* This stage is an optional step and is normally

used in classification tasks such as musical genre or artist classification. Linear Discriminant Analysis (LDA) is applied to minimize intra-class variance and maximize the inter-class variance. The results of this process is reduction of irrelevant dimensions and make the more suitable for classification tasks using a linear classifiers.

5. *Recommendation:* The extracted i-vectors are used in CBRS serving as content features. We should be reminded that here, the UBM is trained on the items in the train dataset and is used an external knowledge source also in the test dataset.

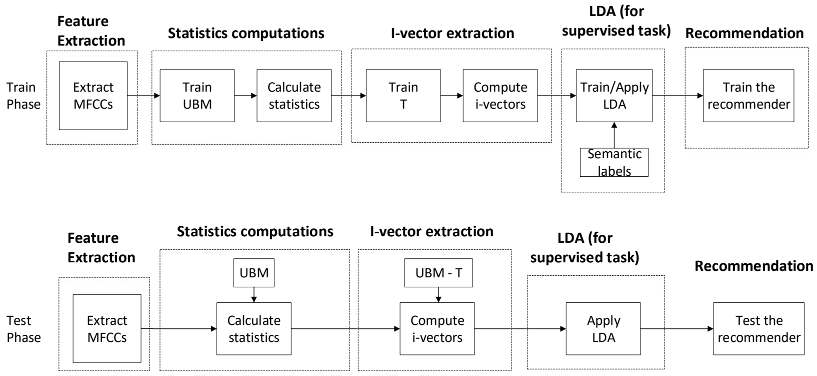


Figure 4.8: *Block diagram of i-vector FA pipeline for supervised and unsupervised approach*

4.4.2 Visual Features

The visual features selected for our experiments are classifiable in two major categories: (i) *aesthetic visual features* used for computational aesthetic measurement in images, and (ii) *deep learning features* extracted from layer FC7 of the AlexNet network developed originally for visual object recognition. The two categories of visual features have been employed in numerous domains including but not limited to media interestingness, photographic image aesthetics, object recognition and emotional classification. Therefore, the proposed visual features can be considered as potentially semantical-rich choices for use in a content-based video recommendation system. We experimented the effect of several feature temporal aggregation methods in order to transform the frame-level features into a video-level descriptor.

Aesthetic-visual features (AVF): this set of visual features has been proposed in [135] for measuring the beauty of coral reefs. Parts of these feature set are inspired from works dealing with artwork, photographic image aesthetics and human perception of images [80, 174, 202]. The set encompasses 26 feature types (feature vectors) which are classifiable into three main categories: *color-based descriptors*, *texture-based descriptors*, and *object-based descriptors*, as proposed by [135], adding up to a feature vector of dimensionality 107 for each frame. The three categories of AVF presented in our collection for recommendation purpose were fused with each other by three different approaches: (i) *individual descriptor* for each of the 26 feature types (ii) feature fusion based on *three categories* of color-based, texture-based and object-based (iii) *combined all* feature vectors by concatenation of 26 descriptors. Furthermore, we applied four different feature aggregation techniques based on statistical summarization functions: *mean*, *median*, *mean and variance* and *median and median absolute deviation*.

1. **Color-based descriptors:** The color-related features in AVF are based on perceptual color spaces such as HSL (Hue, Saturation, Lightness) and HSV (Hue, Saturation, Value). The first property captured is the *average channel value* directly measured from the two color spaces. The second property captured is *the colorfulness* quantified by measuring a number of distance metrics from the 64-bin color distribution of the RGB spectrum and one equal reference distribution. Three distance metrics are considered: *Earth Mover's Distance*, *Quadratic Distance* and *standard deviation* between two distributions. The next type of color features are *hue descriptors*. They provide information on amount of hues contained in each pixel such as number of hues available, number of dominant hues for the image and so forth. A set of nine hue models which are previously validated to be appealing to human [225] serve as a reference to compute a distance with the current picture. The next color-based property is the *brightness* descriptor which is extracted by calculating statistics such as average brightness values and brightness contrast across the image. Ultimately, the average *HSV and HSL* values were calculated by taking into consideration the main focus region and the rule of the thirds compositional rule [237].

2. **Texture-based descriptors:** The texture-related features in AVF are of two types, the *edge* and *texture* components. For the former, statistics obtained from the edge distribution on a frame is used whereas for the latter statistics based on texture range and deviation is used. In or-

der to capture the property related to randomness, the *entropy* of each channel of the RGB color space is calculated. In addition, the *wavelet transformation* based on three level Daubechies was applied on each of the channel of HSV space and together with *average wavelet* were used to serve as a different texture representation [81]. Also, inspired by the *low depth of field* rule of photographic composition, a new feature capturing this property was extracted (further information can be found in [80]). A final texture component inspired by the *low depth of field* photographic composition rule was computed [80]. The set of these features results in 6 feature types for the texture category in AVF

3. **Object-based descriptors:** The object-related features in AVF are based on a method proposed in [80] whose main goal is to find the largest segments in an image by calculating a k-means clustering algorithm. Various features are extracted afterwards including: the *area*, *centroids*, values for the *hue, saturation* and *value* channels, *average brightness values*, horizontal and vertical *coordinates*, *mass variance* and *skewness* for the largest segments in the image (considered as the most salient as well). Finally, in the final descriptor the *color* spread and complementarity are also represented in which the latter component measures hue, saturation and brightness *contrast* between the resulting segments. The set of these features results in 11 feature types for the object category in AVF.

As mentioned earlier, the main advantage of the features defined by AVF is that they are heavily human-oriented. In other words, the role of some components in AVF for psychological or aesthetic aspects of visual communication are widely acknowledged/validated. For instance, inspired by the research work [225], the hue model in AVF is quantified by comparing the distance of the hue model of a certain image with the models considered appealing to humans. Another motivation for the usage of AVF is that they incorporate properties which are inline with general rules in photographic style. The impact of such properties in the human aesthetic perception have been previously acknowledged [193].

Deep-learning features: Deep learning is a specific set of techniques from the broader field of machine learning (ML) that is concerned with the study and use of deep artificial neural networks in order to learn structured representations of data. It has seen a dominate and pervasive resurgence in many research domain including computer vision so much so that it can be considered an inseparable part of the computer vision community. In the context of computer vision, deep learning obtained its main strength for

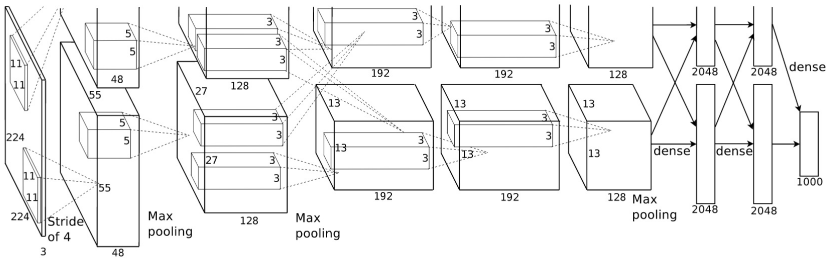


Figure 4.9: *Layers in the AlexNet deep neural network [194]*

the promising visual recognition performance surpassing other methods in artificial intelligence.

The ImageNet Large Scale Visual Recognition Competition (ILSVRC) gave the researcher in the vision community the chance to test various object recognition algorithms on a similar shared dataset. The image dataset represented 1.2 million images taken from ImageNet[4] categorized in $1,000$ different classes. The AlexNet [194] deep neural network was created which was the winner of the competition in 2012 with a significant improvement over the second best method/team in the same year. The authors also ran experiments on the previous edition of ILSVRC 2010 dataset and obtained superior performance compared with SoA approaches of that time.

The architecture of AlexNet deep neural network is shown in Figure 4.9. It contains eight learned layers, five convolutional layers and three fully-connected layers. The network receives as the input a resized $224 \times 224 \times 3$ image. The output of the last fully-connected layer is fed to a 1000 dimensional softmax layer which produces a distribution over the 1000 class labels given each input each. The first convolutional layer filters the $224 \times 224 \times 3$ input image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer takes as input the (response-normalized and pooled) output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 48$. The third, fourth and fifth convolutional layers have the following structure: the third 384 kernels of size $3 \times 3 \times 256$, the forth 384 kernels of size $3 \times 3 \times 192$ and the final convolutional layer has 256 kernels of size $3 \times 3 \times 192$. The fully-connected layers contain $4,096$ neurons each.

Overall the neural network architecture has 60 million parameters and $650,000$ neuron. As described in [194], some data augmentation solutions are employed in order to reduce the chance of overfitting. The data augmen-

[4] http://www.image-net.org/

tation transforms the image into a new image in such a way that it can be produced from the original images with very little computation. The transformations include: image translations, horizontal reflections and altering of the intensity of the RGB channels [152, 194]

As pointed out in [194], the main novelty introduced by this network is on the introduction ReLU (Rectified Linear Units) nonlinearity output function defined by $f(x) = max(0, x)$ which allows faster training times compared with conventional functions such as $f(x) = tanh(x)$ or $f(x) = (1 + e^{-x})^{-1}$.

Given the promising performance in other tasks/fields, we extract fc7 features for each frame. Therefore, the final descriptor has the same dimensionality as the output softmax layer which is equal to 4096. Similar to previously-described AVF, frame-level descriptors are transformed into video-level descriptor by using statistical summarization function such as *mean*, *median*, *mean and variance* and *median and median absolute deviation*. Additionally, we tested vector of linearly agregated descriptors (VLAD) aggregation method using PCA for dimensionality reduction, with three different sizes for the visual word codebook: $k \in \{32, 64, 128\}$

4.5 Evaluation Setup

In this section, we describe the details about the experimental setup for the validation of the proposed recommendation framework. In particular, we discuss the content-based descriptors used for the experiments (multimedia and metadata), the recommendation algorithm, our proposed rank-based hybridization algorithm for multimedia integration and parameter tuning.

4.5.1 Content Descriptors

In both studies (the offline experiment and the user-study) as shall be presented in the following, the quality of recommendation has been evaluated using an identical CB recommendation algorithm (see section 4.5.2) and different content descriptors serving as the manipulatable variable.

Multimedia Features

The proposed multimedia features can be classified based on the period of time they have been exploited in the community[5]. From this perspective, we can classify the multimedia features in two major groups:

[5] Mainly the communities of computer vision, multimedia and music information information retrieval.

1. **The established traditional features**: The features in this category include (i) block-level audio features (BLF) and, ii) aesthetic visual features (AVF).
2. **The novel state-of-the-art features**: The features in this category include iii) the i-vector audio features and, (iv) AlexNet deep visual features.

As for the best temporal aggregation method for the visual category, between two categories of aggregation methods we tried, *i.e.,* statistical summarization and VLAD, it was observed that statistical summarization methods overall provided substantially better performance compared with VLAD. It was also noted that the simple *mean* statistical summarization function provided relatively better performance compared with the rest of functions *e.g., median*. We therefore chose the *mean* statistical summarization function to compare and report the quality of different recommendation approaches.

Regarding the parameters associated with i-vectors, we extracted 20-dimensional MFCCs from the items in the train dataset. They were used to train a UBM with number of Gaussian components equal to 256 or 512. We tried different dimensionalities of the latent factor variable (the length of the final i-vector feature) equal to $(40, 100, 200, 400)$. We performed a hyper-parameter search by trying each of the $2 \times 4 = 8$ possible combinations. We chose the best performing parameters with regards to the under-study metric and recommendation-list size based on the average performance obtained over 5-fold cross validation.

Metadata attributes

Two types of metadata features are used in our system: (i) *genre features* (editorial metadata) (ii) *tag Features* (user-generated metadata).

The genre feature contains 18 genre labels, namely: *action*, *adventure*, *animation*, *children*, *comedy*, *crime*, *documentary*, *drama*, *fantasy*, *film-Noir*, *horror*, *musical*, *mystery*, *romance*, *sci-Fi*, *thriller*, *war*, and *western*. The final genre feature vector is a binary vector of dimensionality 18; The tag feature is based on TF-IDF Bag-of-Words (BoW) model. Beforehand, a series of preprocessing steps are adopted to prepare the textual tag features into usable numerical feature vectors. The steps include: (i) punctuation removal, (ii) tokenization and lowercase conversion, (iii) stop-word removal, (iv) very short or very long word removal which means we remove words with 2 or fewer characters, and words with 15 or greater, and (v) porter stemming, the process of reducing inflected/derived words to their stem or root form [252]. Ultimately, a BoW model based on term frequency inverse document frequency (TF-IDF) is created which can represent how important a word is to a document *w.r.t.* a collection of documents. We used the "word counts" for the TF term and the logarithmic variant $\log(N/N_t)$ for the IDF term in which N_t is the number of documents containing the word t and N is the total number of documents in the collection.

The RS community for long has used metadata as the main and only type of content description in CB or Hybrid recommendation systems since they are human-generated and are semantically rich. We therefore use metadata features as a baseline during the evaluation. We observed that the tag features constantly outperformed the genre. Therefore, tag features will represent the main baseline for comparison. In addition, we use metadata in conjunction with the multimedia descriptors to build hybrid RS that provides higher quality of recommendations compared to the RS using each of the features alone.

4.5.2 Recommendation Algorithm

Recommendations are generated by using a widely used pure content-based filtering (CBF) algorithm based on k-nearest neighbors. The unknown rating (preference score) $\hat{\mathbf{r}}_{ui}$ for user u and item i is calculated as a weighted average score of the ratings of the most similar items defined by their the CB similarity as given by Eq 4.2

$$\hat{\mathbf{r}}_{ui} = \frac{1}{\sum_{j \in N_u(i)} s_{ij}} \sum_{j \in N_u(i)} s_{ij}\, \mathbf{r}_{ui} \tag{4.2}$$

in which $N_u(i)$ denotes the items rated by user u which are most similar to item i and s_{ij} is the similarity score between items i and j (the CB similarity) [112]. For all the content descriptors we used 10 neighbors and cosine similarity metric. This is with the exception of genre feature for which we also tried Jaccard similarity metric, since that Jaccard similarity metric suits better for binary features. It was observed that the Jaccard similarity provided slightly better performance compared with cosine similarity metric and was chosen to report the final results for genre feature.

4.5.3 Hybridization

We propose a rank-based hybrid recommendation method to enhance the performance of individual CB systems. The method is the extended version of the Borda count rank aggregation method which was used successfully in a similar domain, like group recommendation [34]. The Borda count method is a rank aggregation method which awards each item in the recommender's rank list with a score in accordance with its rank position in the list (the lower the position, the higher the score).

Given two personalized ranking lists $\sigma_u^{rec_a}$ and $\sigma_u^{rec_b}$ produced be recommenders a and b for user u, the combined score of item i for user u based on Bodra count is given by:

$$\mathrm{CS}(u,i) = \frac{N_a - \sigma_u^{rec_a}(i) + 1}{N_a} + \frac{N_b - \sigma_u^{rec_b}(i) + 1}{N_b} \tag{4.3}$$

where N_a and N_b are the total numbers of items in the corresponding rankings. As it can be seen, only in the case where an item stays consistently on the top of the recommendation list (*i.e.,* $\sigma_u^{rec_a}$ and $\sigma_u^{rec_b}$ are both small), the combined score for that item would be large. The main shortcoming of this approach is that if one voter (recommender) provides good-quality rankings in such a way that it dominates the performance, the original Bodra count method using equal weights to both voters, can push the overall results toward the middle-point quality of the combined voters resulting in lower qualities when the differences between two recommendation qualities are large.

To this end, in order to enhance more the performance of the system, we propose to extend the standard Borda count aggregation method given by Eq. 4.3 using an aggregation approach based on *weighted averaging*, yielding

$$\mathrm{CS}(u,i) = \sum_{k=1}^{k=V} w_k \cdot \frac{N_a - \sigma_u^{rec_k}(i) + 1}{N_a} \tag{4.4}$$

where V is the number of voters and w_k is the importance weight associated with voter k. The advantage of the proposed hybridization approach is that it can make a fair balance between the voters and improve it with the complementary information of the other voter(s) leading to improved overall performance.

In this research work, we present the results of hybridization only for the offline study. The weight values lie in the range $[0, 1]$ in such a way that $w_1 + w_2 = 1$. We tried 100 combinations in a linear span of $[0, 1]$ and found the best weights *w.r.t.* each evaluation metric and list size based on the average performance obtained over 5-fold cross validation. It was interesting to note that quality of the the proposed hybrid system significantly outperforms the standard Borda count method when the search space grows (*e.g.,* from 10 to 50 and to 100) [34].

Table 4.1: *Characteristics of the evaluation dataset used in the offline study: $|\mathcal{U}|$ — number of users, $|\mathcal{I}|$ — number of items, $|\mathcal{R}|$ — number of ratings.*

dataset	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	$\frac{\lvert\mathcal{R}\rvert}{\lvert\mathcal{U}\rvert}$	$\frac{\lvert\mathcal{R}\rvert}{\lvert\mathcal{I}\rvert}$	$\frac{\lvert\mathcal{R}\rvert}{\lvert\mathcal{I}\rvert\times\lvert\mathcal{U}\rvert}$ (density)
ML-20M	3,000	4,899	212,019	70.67	43.27	0.0144

4.6 Experimental Study A: Offline Experiment with Simulated Users

In this section, we explain the offline experiment which uses historical dataset. The offline experiment is performed to simulate the cold- and warm- start scenarios. The specific experimental setup is presented in the following.

4.6.1 Data

We used the MovieLens-20M (ml-20m) dataset [142] which provides timestamped user–item interactions with an up-and-running RS in order to facilitate research on personalized movie recommendation. In order to speed up the experiments, we randomly select a subset of users from the ml-20m each having a minimum of 50 ratings in their rating profile. The characteristics of the final rating dataset is shown in Table 5.3. We used 5-fold cross validation (CV) to perform all the offline experiments. This is realized by partitioning the items in our dataset into 5 non-overlapping subsets (item-wise splitting of the user-rating matrix). We intentionally chose item-wise splitting of the rating dataset in order to step our attention aside from the discussion of CF systems, since CF are not usable in such splitting system when items in the dataset do not have any ratings (only CB systems can be used) (see section 4.9).

4.6.2 Objective Evaluation Metrics

The evaluation metrics used for the offline experiment are of two natures: (1) *accuracy metrics* which measure the relevance of the recommendations *w.r.t.* users' preferences such as recall and mean average precision and, (2) *beyond-accuracy metrics:* which focus on others aspect of user satisfaction beyond relevance such as diversity, novelty and catalog coverage. Some of the latter metrics types have a direct impact on business profits. In the following we formally define the evaluation metrics adopted in this offline experiments:

(1) Accuracy Metrics. We used the following evaluation metrics as listed below:

- *Mean average precision (MAP)* is a metric that computes the overall precision of a recommender system based on precision at different recall levels [205]. It is computed as the arithmetic mean of the average precision (AP) over the entire set of users in the test set, where AP is defined as follows:

  $$AP = \frac{1}{\min(M,N)} \sum_{k=1}^{N} P@k \cdot rel(k) \tag{4.5}$$

  where $rel(k)$ is an indicator signaling if the k^{th} recommended item is relevant, i.e. $rel(k) = 1$, or not, i.e. $rel(k) = 0$; M is the number of relevant items and N is the number of recommended items in the top N recommendation list. Note that AP implicitly incorporates recall, because it considers relevant items not in the recommendation list. Finally, considering the AP equation, MAP will be defined as follows:

  $$MAP = \frac{1}{|U|} \sum_{u \in |U|} AP_u \tag{4.6}$$

  where U denotes the users in test set.

- *Mean reciprocal rank* is a metric that evaluates the results of a recommender system based on the order of probability of correctness. It is defined as following:

  $$MRR = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{rank_u} \tag{4.7}$$

  where $rank_u$ is the position of the first relevant retrieved answer for user u. If no relevant answers are retrieved, the value will be 0.

- *Recall at top K recommendations (R@K)* is presented here for the sake of completeness, even though it is not a crucial measure from a consumer's perspective. Indeed, the user is typically not interested in being recommended all or a large number of relevant items, rather in having good recommendations at the top of the recommendation list. R@K is defined as:

  $$R@K = \frac{1}{|U|} \sum_{u \in U} \frac{|L_u \cap \hat{L}_u|}{|L_u|} \tag{4.8}$$

where L_u is a set of relevant items of user u in the test set T and $\hat{L}_u$ denotes the recommended set containing the K items in T with the highest predicted ratings for the user u from the set of all users U.

(2) Beyond Accuracy Metrics. Coverage of a recommender system is defined as the proportion of items over which the system is capable of generating recommendations [151]:

$$coverage = \frac{|\hat{T}|}{|T|} \tag{4.9}$$

where $|T|$ is the size of the test set and $|\hat{T}|$ is the number of ratings in T for which the system can predict a value. This is particularly important in cold start situations, when recommender systems are not able to accurately predict the ratings of new users or new items, and hence obtain low coverage. Recommender systems with lower coverage are therefore limited in the number of items they can recommend. A simple remedy to improve low coverage is to implement some default recommendation strategy for an unknown user–item entry. For example, we can consider the average rating of users for an item as an estimate of its rating. This may come at the price of accuracy and therefore the trade-off between coverage and accuracy needs to be considered in the evaluation process [22].

Novelty measures the ability of a recommender system to recommend new items that the user did not know about before [15]. A recommendation list may be accurate, but if it contains a lot of items that are not novel to a user, it is not necessarily a useful list [349]. While novelty should be defined on an individual user level, considering the actual freshness of the recommended items, it is common to use the self-information of the recommended items relative to their global popularity:

$$novelty = \frac{1}{|U|}\sum_{u \in U}\sum_{i \in L_u}\frac{-\log_2 pop_i}{N} \tag{4.10}$$

where pop_i is the popularity of item i measured as percentage of users who rated i, L_u is the top-N recommendations for user u [349, 353]. The above definition assumes that the likelihood of the user selecting a previously unknown item is proportional to its global popularity and is used as an approximation of novelty. In order to obtain more accurate information about novelty or freshness, explicit user feedback is needed, in particular since the user might have listened to an item through other channels before. It is often assumed that the users preferred recommendation lists with more novel items. However, if the presented items are too novel, then the user is unlikely to have any knowledge of them, nor to be able to understand or rate them. Therefore, moderate values indicate better performances [178].

Diversity is another important beyond-accuracy measure which gauges the extent to which recommended items are different from each other, where difference can relate to various aspects, *e.g.,* musical style, artist, lyrics, or instrumentation, just to name a few. Diversity can be defined in several ways. One of the most common is to compute pairwise distance between all items in the recommendation set, either averaged [354] or summed [300]. In the former case, the diversity of a recommendation list L is calculated as follows:

$$diversity(L) = \frac{\sum\limits_{i \in L} \sum\limits_{j \in L \setminus i} dist_{i,j}}{|L| \cdot (|L| - 1)} \tag{4.11}$$

where $dist_{i,j}$ is the some distance function defined between items i and j. Common choices are inverse cosine similarity [267], inverse Pearson correlation [318], or Hamming distance [175].

(3) Different Cut-off Values. Cut-off value is the size of recommendation list returned to the user. We use two cut-off values to compute recommendation qualities at. They include Rec@4 and Rec@10. Rec@10 is a common choice in many works. Morover, given the average ratio of 70.67 ratings per user and 5-fold CV by splitting the user-rating matrix item-wise, on average, 14 ratings exists for each item in the test dataset which makes the recommendation at 10 effective. Regarding the former, Rec@4 was chosen in order to provide the possibility of comparing offline results with the result of user-study (see next section) since all recommendations are computed at single cut-off value of 4 in the user-study. The reason for this choice is motivated in the next section.

4.6.3 Recommendation *w.r.t.* Accuracy in Cold-Start Scenario

In the offline experiment, a cold-start (CS) scenario[6] is simulated which concerns solely about tag features. Tags are user-generated keywords. Often around the time a new item is added to the catalog, the item can have no or very few associated tags. We can simulate a CS scenario for the user-generated tag features by sampling a small percentage of the tags for the

[6] As reminded earlier, the CS scenario addressed here is the metadata CS, where the items lack sufficient metadata. We intentionally step our attention aside from rating CS.

Table 4.2: *Performance of various audio-visual feature sets w.r.t. Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and Recall in the* ***Cold Start*** *scenario. All results are based on a 5-fold cross-validation. The first and second best results are shown in* ***red*** *and* ***blue****, respectively.*

			@4			@10		
	feature name	feature type	MRR	MAP	Recall	MRR	MAP	Recall
Unimodal	tag	metadata	0.0195	0.0051	0.0042	0.0274	0.0037	0.0111
	genre	metadata	0.0162	0.0044	0.0039	0.0245	0.0034	0.0112
	i-vector (SoA)	audio signal	0.0233	0.0060	0.0052	0.0311	0.0042	0.0120
	BLF (traditional)	audio signal	0.0170	0.0045	0.0038	0.0242	0.0032	0.0097
	Deep (SoA)	visual signal	0.0219	0.0057	0.0043	0.0296	0.0038	0.0111
	AVF (traditional)	visual signal	0.0187	0.0049	0.0039	0.0263	0.0034	0.0102
Hybrid	i-vector + Deep	audio + visual	0.0232	0.0061	0.0051	0.0318	0.0043	0.0122
	Deep + tag	visual + meta	**0.0239**	**0.0062**	**0.0053**	**0.0325**	**0.0044**	**0.0130**
	i-vector + tag	audio + meta	**0.0266**	**0.0072**	**0.0059**	**0.0359**	**0.0049**	**0.0139**

items in the test dataset while discarding the rest of tags. We are required however to use the complete tags for the items in the train set since we can always assume our recommender system is trained with a mature and complete set of tags. The challenge is to see how it can cope with CS situation manifest in sparsity of tags. To this end, we kept 3% of the tags associated with the items in the test selected in a random fashion and discarded the tags for the rest of items. In section 4.6.4 however, we present the results for the warm-start (WS) scenario to provide a realistic picture of the overall system performance in different scenarios. In both CS and WS, the genre features are assumed to be accessible as metadata (genre feature are released together with the release of a movie), therefore we use complete genre labels in these scenarios.

In Table 4.2, we present the results of the offline experiment for the CS simulation phase. Two types of CB recommenders are considered: unimodal and multimodal depending on if one or more than one modality (audio, visual and textual) is used on the content description of the systems. The multimodal recommender are built by using the proposed rank-based hybidization apporoach described in section 4.5.3. All the results are presented with regards to three different accuracy metrics namely MRR, MAP and Recall with cutoff values @4 and @10. As expected, the results are consistent across all presented metrics for the top recommenders shown in **red** and **blue**, respectively. Finally, as stated earlier, tags has the best value from the metadata features collection, and therefore it serves serves as the main baseline for comparing the results.

It can be noted that the SoA audio (i-vector) and SoA visual (Deep) features outperform metadata in unimodal recommendation *w.r.t.* MRR, MAP and Recall with regard to all test cut-off values @4 and @10. The difference between SoA audio and tags however is relatively larger than the one between SoA visual and tags. More specifically, in the @4 cutoff experiments, both i-vector and Deep features outperform the tag baseline considerably, raising the scores for MRR, MAP and Recall by 19.4%, 17.6% and 23.8% respectively (in the i-vector case) and by 12.3%, 11.7% and 2.3% (in the Deep feature case). In the @10 cutoff experiments, i-vector consistently provides a good degree of improvement by 13.5%, 13.5% and 8.1%. The Deep visual feature show a lower improvement in the MRR and MAP metric by 8% and 2.7 % while pretty similar scores are obtained for Recall. In addition, it can be noted as well that the traditional features (BLF audio and AVF visual) have lower scores compared with the tag feature in almost all experiments signifying the informativeness of the SoA features for recommendation task. We chose the best performing candidates from the unimodal phase (*i.e.,* i-vector, Deep and tag features) as the candidates for the multimodal combination schemes.

Combining the advantages of multimedia and metadata using our proposed hybridization via extended Borda rank aggregation method leads to improvement of the overall performance for the majority of combinations (in CS and WS) substantially or significantly. For instance, in CS the i-vector + deep hybridization method improves the tag performance by 18.9%, 19.6% and 21.4%. Although in some cases, the latter difference is comparable with one of the two unimodal components, the improvement clearly indicates different type of information encoded by multimedia signals not directly reflected in the metadata. In this regard, it can be observed that the i-vector + tag combination significantly improves the metric scores in the @4 cutoff experiments by 36.4%, 41.1% and 40.4% while the improvement for Deep + tag still remains high equal to 22.5%, 21.5% and 26.1%. As for the @10 experiments, the i-vector + tag combination shows an improvement of 31.0%, 32.4% and 25.2% *w.r.t.* tag while the same metric for Deep + tag is 18.6%, 18.9% and 17.1%.

These results are interesting and are indicative of the fact that in CS scenario SoA multimedia features, even in their unimodal form, represent a useful approach for recommendation generating a superior quality for most of the metrics and most of the cutoff values. In addition, these results show the complementary information encoded in the actual multimedia features (mainly Deep visual and i-vector) not directly captured by the user-generated tag features.

4.6.4 Recommendation *w.r.t* Accuracy in Warm-Start Scenario

In warm start (WS) scenario, all the metadata features are available (specifically speaking the tag features). This would lead to obvious improvement in the quality scores obtained by the tag feature in CS and WS scenarios, for example compare MRR@4 in two cases: 0.0195 (for CS) and 0.0213 (for WS). The results for both unimodal and multimodal hybrid approaches are presented in Table 4.3 which in a similar fashion are presented with regards to MRR, MAP and Recall metrics with @4 and @10 cutoff values. The immediate observation is that ranking of these features and their combinations in terms of their performance is consistent throughout all these tests. The two feature combinations Deep+tag and i-vector+tag are consistently the best compared with all six unimodal cases as well as the i-vector+Deep multimodal case.

Table 4.3: *Performance of various audio-visual feature sets w.r.t. Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and Recall in the* ***Warm Start*** *scenario. All results are based on a 5-fold cross-validation. The first and second best results are shown in* ***red*** *and* ***blue****, respectively.*

			@4			@10		
	feature name	feature type	MRR	MAP	Recall	MRR	MAP	Recall
Unimodal	tag	metadata	0.0213	0.0057	0.0046	0.0294	0.0041	0.0120
	genre	metadata	0.0162	0.0044	0.0039	0.0245	0.0034	0.0112
	i-vector (SoA)	audio signal	0.0233	0.0060	0.0052	0.0311	0.0042	0.0120
	BLF (traditional)	audio signal	0.0170	0.0045	0.0038	0.0242	0.0032	0.0097
	Deep (SoA)	visual signal	0.0219	0.0057	0.0043	0.0296	0.0038	0.0111
	AVF (traditional)	visual signal	0.0187	0.0049	0.0039	0.0263	0.0034	0.0102
Hybrid	i-vector + Deep	audio + visual	0.0232	0.0061	0.0051	0.0318	0.0043	0.0122
	Deep + tag	visual + meta	**0.0267**	**0.0066**	**0.0054**	**0.0358**	**0.0047**	**0.0139**
	i-vector + tag	audio + meta	**0.0295**	**0.0082**	**0.0067**	**0.0400**	**0.0056**	**0.0156**

Based on these results one can note that the unimodal approaches based on SoA audio and visual features beat the traditional features by large margins, for instance consider the audio features i-vector (SoA) *v.s.* BLF (traditional) scores for MRR@4 equal to 0.0233 to 0.0170 (37 % higher) and the same for the visual features Deep (SoA) 0.0219 to 0.0187 (17 % higher). As for other features, it can be seen that i-vectors outperforms the best performing metadata feature, tag at diferent chosen cutoff values except Recall@10, where the two have a similar performance. The best percentual performance growth for i-vectors over the tag baseline are equal to 13% for Recall@4 (0.0052 *vs* 0.0046) and 9.3% for MRR@4 (0.0233 *vs* 0.0213). The deep features can provide better quality scores over tags only for MRR and perform equally or slightly worst compared with tags for all metrics and cut-off values. This result indicates that in WS situation if the goal of the RS is to provide *single* good recommendation to the users, the visual feature can be safely and better replaced with tags. In absence of this condition, tags features are more advised to be used!

Similar to CS, the three hybridization methods chosen for WS are the three winners of the unimodal case: i-vectors, AlexNet Deep and tag. Interestingly, the observation regarding the complementariness of i-vector + Deep in comparison with tag remains the same in the WS scenario. Moreover, the multimodal combinations Deep + tag and i-vector + tag outperform the unimodal components by a significant margin. For instance in cut-off value @4, the i-vector + tag combination outperforms tag by 38.5%, 43.8% and 45.6% which is a great improvement. The second combination Deep + tag outperform tag by 25.3%, 15.7% and 17.3% which is as well quite substantial.

The results in both CS and WS scenarios agree that the multimedia features performed best when fused with tag features therefore confirming that these different types of features describe the video content in a complementary fashion and can meet users' differing information needs. These results are novel and opposed to the general view in the RS community which considers metadata as complete semantically rich features. The promising performance obtained by the SoA features was one of our expectations regarding the development of this system, considering that these features are used in various related domains related to recommender systems such as music and speech and acoustic scene modeling for the i-vectors and emotion recognition, interestingness and aesthetics for the visual features. Ultimately, the consistent scores obtained across different metrics used, suggest the reproducibility of the results across different approaches.

4.6.5 Recommendation *w.r.t* Beyond Accuracy Metrics

In this section, we focous on the performance of different feature *w.r.t.* different Beyond Accuracy Metrics. These metrics capture properties beyond the relevance of the content with the user profile which are equally important in overall user satisfaction [22, 133]. Some of these metrics have an immediate impact on business revenue (*e.g.,* catalog coverage or novelty). We should heed to the fact that since the final objective of a designer is to select a recommender that provides maximum performance *w.r.t.* accuracy and beyond accuracy metric(s), practically speaking it is not possible to enter the worst recommender from previous section to the features comparison in this section, mainly because maximizing the beyond-accuracy metric for a particular feature makes only sense when we have obtained a minimum acceptable performance with regards to accuracy. For example, we can mention a random recommender as the recommender type providing the most diverse recommendation. However, it is obvious that random this recommender cannot not provide high quality recommendations since they are not personalized for each user. Speaking about diversity maximization makes sense only when the recommender passes a minimum threshold for accuracy score and maintain a balance between the two.

Therefore, in this section, we base the discussion only on the “winner” features/recommenders from the accuracy section (mainly WS scenario). Specifically, we do not focus on the results of genre recommender since it has the lowest performance among all feature sets by large margins. Instead, our main goal is to compare the quality of the hybrid recommenders Deep + tag and i-vector + tag with tag as the only baseline. For completeness however, we show the results for genre in gray color.

Table 4.4: *Performance of various audio-visual feature sets w.r.t. Novelty, Diversity, and Coverage in a* ***Warm Start*** *scenario. All results are based on a 5-fold cross-validation. The first and second best results are shown in* ***red*** *and* ***blue****, respectively. Please note that given the poor performance of genre for accuracy metrics, we use tag as the main baseline and show genre in gray, thus excluding it's results from calculating the first two best results, however keep the results of genre for completeness.*

			@4			@10		
	feature name	feature type	Novelty	Diversity	Cover-age	Novelty	Diversity	Coverage
Unimodal	tag	metadata	**8.9845**	0.8685	0.9700	**8.8860**	0.8681	0.9992
	genre	metadata	9.1759	0.7553	0.9383	8.9974	0.7733	0.9990
	i-vector (SoA)	audio signal	8.7934	**0.8747**	0.9994	8.7565	**0.8756**	**1.0000**
	BLF (traditional)	audio signal	8.6528	0.8618	**0.9998**	8.6580	0.8624	**1.0000**
	Deep (SoA)	visual	8.6397	0.8665	**1.0000**	8.6222	0.8670	**1.0000**
	AVF (traditional)	visual	8.6346	0.8735	0.9994	8.6354	0.8739	**1.0000**
Hybrid	i-vector + Deep	audio+visual	8.7848	**0.8750**	**1.0000**	8.7588	**0.8758**	**1.0000**
	Deep + tag	visual+meta	**8.9720**	0.8671	**1.0000**	8.8824	0.8679	**1.0000**
	i-vector + tag	audio+meta	8.9699	0.8727	0.9994	**8.8880**	0.8751	**1.0000**

Novelty. From the results presented in Table 4.4, it is interesting to understand that both type of MM features (audio and visual) produce the least novel recommendation. This is while tag or tag-hybrid based approaches (*e.g.,* deep+tag or i-vector+tag) provide the most novel recommendations. This is excluding the genre recommender which is capable of exhibiting the highest degree of novelty among all recommendation techniques. We think the explanation for this performance is that genre features are not sensitive to popularity of movies which according to Equation 4.10 is inversely proportional to novelty (novelty equals unpopularity). In other words, from a genre-recommender perspective, two movies are similar *w.r.t.* a reference movie as long they belong to the same genre, irrespective of their popularity. Despite this result, as mentioned earlier we step our attention aside from the genre recommender due to its low performance with regard to accuracy metrics. On other hand, as for the performance of MM features we can note previous works [103] in which multimedia features shown a high degree of correlation with popularity and thus it is likely that popular movies remain closely clustered in the feature space defined by multimedia features. In [103], the authors successfully included a number of low-level and high-level visual features based on aesthetic and compositional measures in order to predict social image interestingness defined by Flickr's interestingness score. The latter score is a measure based on the popularity of the image or creator, viewing patterns and so forth. We therefore think since most of the users have popular movies in their rating profiles, given a target user profile a CBRS based on multimedia content will have a higher chance to retrieve popular movies compared with genre-based CBRS. This is motivated based on the hypothesis that multimedia features and popularity show a higher degree of correlation compared with genre features and popularity.

From the hybridization viewpoint, as it can be seen the tag feature outperforms the multimedia features, with tag's scores of 8.9845 for Novelty@4. This is while Deep+tag provides the second best performance with the score of 0.89720. As for Novelty@10, the ranking is inversed and the i-vector+tag provides the highest novel recommendation with the score of 8.8880 while tag scores 8.8860. The conclusion that can be drawn is that "*tag features promote novelty either as a standalone feature or in conjunction with MM features*".

Another important observation is that the gain margins between the minimum and maximum values for novelty is quite marginal equal to 6.2% for Novelty@4 and even lower, 4.3% for Novelty@10. Therefore, there exists an insignificant trade-off regarding novelty compared to the great improvement in accuracy 20% to 40% obtained by the hybrid techniques when taking into account the recommender metrics presented in Subsections 4.6.3 and 4.6.4.

Diversity. In order to compute diversity, we need to define a property based on which diversification is defined. While different candidate are available such as genre, tag and MM features, in this experiment, we choose diversification based on genre since first it is common practice to use genre as a diversification measure given that genre features are available with movies and second that genre remains outside the competition between features in this part of the experiment and is a fair property for computing diversity (*e.g.,* consider that a recommender based on tag would be less likely to promote diversity defined based on tag). We calculate diversity by computing pairwise intra-list distances between recommended movies based on genre similarity and the jaccard similarity metric (Dist = 1- Sim).

As for the diversity result, all features except genre achieved scores between 0.8618 and 0.8750 for Diversity@4 and 0.8624 and 0.8758 for Diversity@10 leaving a very small margin between the maximum and minimum score with regards to diversity. MM features exhibit a different behavior with regards to diversifying recommendations in which the SoA audio (i-vector features) and the traditional AVF features both outperform metadata features in the unimodal scenario. The most diverse recommender for both cut-off values Diversity@4 and Diversity@10 are i-vector+deep which again is an expected result, since these two components are not metadata features and build a differently distributed feature space when compared to tag and genre. in this regard, when any of the best-performing SoA multimedia features are fused with the tag feature they do not reach the maximum score, however they score close to i-vector + deep fusion. As mentioned earlier, the difference margin for diversity at both cut-off values is as small as a 1.5%. Therefore, there exists is a proper trade-off between accuracy and diversity when keeping in mind the great advantages brought by the proposed hybridization technique with regards to accuracy measure.

Coverage. As mentioned earlier, catalog coverage is an important metric, especially for a CS scenario. In our experiments, it was observed that the maximum value of 1 was scored by two hybrid approaches (i-vector + deep and deep + tag) and unimodal system (deep), even at such a low cutoff value. This can be indicative of the fact that visual features ensure a maximum coverage on the dataset. In this regard, the performance of audio features remain very close (0.9994 for both i-vector and i-vector + tag hybrid approach) also creating a good RS from a coverage perspective. It is interesting to see that tag feature have the worst performance *w.r.t* coverage. As for, Coverage@10 all recommenders that contain a MM component provide a maximum score while tag showing the lowest diversity score (0.9992).

Based on these results the conclusion that can be drawn is that "*the best approaches with regards to coverage are the ones that contain audio or visual features. This can be an evidence that MM features can guarantee* *higher degrees of coverage when used unimodal or hybridized with metadata features.*".

4.7 Experimental Study B: Online Evaluation

In this section, we present an empirical user-study conducted with the use of a developed working movie recommender system similar to Netflix. The system is an extension of *MISRec* presented in section 3.3.4; it uses the same CBRS algorithms described in the previous offline experiment powered by visual, audio and metadata content descriptors as the manipulatable variable. The goal of the system is to measure user's perceived quality of the recommendation with regards to: *accuracy*, *novelty*, *diversity*, *understanding the user* and *overall satisfaction*.

We intentionally do not use the proposed hybridization technique in the offline experiment on *MISRec* recommendation engine in order not to overload the user with numerous selection choices that she can have among different recommendation techniques with the aim to obtain more reliable feedbacks. Therefore, the main recommenders considered in the Experimental Study B are only the unimodal recommendation schemes: (i) metadata: *genre* and *tag*, (ii) audio: *i-vector* and *BLF*, and (iii) visual: *Deep-learning features* and *AVF*.

4.7.1 Perceived Quality Metrics

Perceived quality is a vague notion. It can be considered as an indirect indicator of a recommender's potential for persuasion [76]. Perceived quality is defined as the extent to which users judge recommendations positively and value the general experience with the recommender system. In this study, the notion of perceived quality is operationalized along five dimensions [110]:

1. *Perceived accuracy* (aka *Relevance*): It measures the degree of match between recommendations and the users' preferences;

2. *Satisfaction*: It measures the overall users' feeling of engagement with the RS;

3. *Understands Me*: It measures the perceived personalization or the user's perception that the recommender comprehends their preferences and can effectively adjust to them;

4. *Novelty*: It measures the degree to which users receive new recommended movies;

5. *Diversity*: It measures how much users perceive recommendations as dissimilar in relation to each other, *e.g.* movies from different genres.

We used a survey containing 22 questions originally taken from [182] with a specific end goal to quantify the user's perception of the recommendation lists across five factors. As suggested in [110] the questions are posed in comparative manner [110] instead of absolute value by asking the users to choose one list among three for each questions of the five dimensions.

4.7.2 Evaluation Protocol

A web-centric testing framework is designed to facilitate the execution of a controlled empirical study for the movie domain as shown in Figure 4.10. The system is named *MISRec* and is powered by the same CBRS algorithms used in Experimental Study A: Offline Experiment with Simulated Users. For instance, MISRec supports its users with a broad range of functionalities that are provided in online video on demand services such as Netflix (`https://www.netflix.com/`) and Lovefilm. As mentioned in Chapter 3 - study 2, MISRec contains the same catalog the movies used in the Offline experiment. Users can explore the catalog, obtain detailed description about each movie and receive recommendations. Additionally, it embeds a survey containing standard questionnaire in order to measure the perceived quality across each of five dimensions: *accuracy*, *novelty*, *diversity*, *understanding the user* and *overall satisfaction*. The questionnaire would allow the designer to collect qualitative information from the users about the system in a relatively easy manner. It is worthwhile to remind that the first prototype of MISRec used for an empirical study on the movie domain was published at ACM conference of Recommender Systems in [112].

The main research subjects have an age range between 18 to 50 years. They have basic understanding of the Web but never used MISrec before the study. This is important to account for likely biases or misunderstandings that can occur from the previous uses of the system. The total number of recruited users was 82 in which 70% were male and 30% female. The mean and the standard deviation of users' ages are 23.93 and 5.11 years.

The entire process of user interaction with the system is illustrated in Figure 4.10. In the first step, the user is asked to sign up by providing her email address, user name and password (Figure 4.10 top middle). The users are taught that they are free to provide an anonymous Email address like whatever_you_like@anonymous.mes if they prefer not to provide their actual Email addresses. In the next step, the user is invited to specify basic information about herself such as her demographic information: age, sexuality, level of education, nationality and other information specific to movie domain: number of movies watched per month, the consumption channel and some optional social network Ids including Twitter, Facebook and Instagram. Then, the user is guided to a personality assessment page where she is required to fill out the Ten-Item Personality Inventory (TIPI) questionnaire (Figure 4.10 middle-left) so that the system can measure her big five personality traits. The information obtained to this point account for user's explicit specification of information about herself. In the next step, the goal is to obtain her preference about movies in an implicit manner via conducting a preference elicitation [67] in which the user is invited to explore the movie catalog by selecting her favorite genre and navigating through different years of production and choose four of her favorite movie. The user can easily select her favorite movie, watch the trailers of the selected four movies online and provide her preference toward each movie explicitly via a 5-level Likert scale rating (1 = low interest for/appreciation of the movie; to 5 = high interest for/appreciation of the movie). The entire process is done in a user-friendly manner and is conducted with the goal of building a user profile representing user's tastes about movies (the user profile is built *w.r.t.* different content features). If for an unknown reason, a movie is not displayable, the user can report the movie and it would be automatically skipped and excluded from the user profile generation. Finally, on the basis of the ratings collected and content descriptors described in Section 4.4, three recommendation lists are presented to the user.

We decided to provide four recommendation per list (Rec@4) because watching the trailers can be time consuming. For a similar reason, we decided not present the user with six recommendation lists using any of these features since it can simply bore the user. This happens in *between-subject design* where each subject has to test all variants of the recommendations techniques. Instead, we chose a *between-subject design* where recommender candidates are randomized for a given subject. As our main objective is to have each user compare **three** CBRS techniques based on: audio, visual and metadata at the same time. Therefore, the way we implemented the between-subject design contemplated of randomizing two instances of each class for a given user, *i.e.,* (BLF or i-vector) for audio, (AVF or Deep) for visual and (genre or tag) for metadata. This results in a user comparing one out of eight possible combinations at a time *e.g.,* (BLF, AVF, genre), (i-vector, AVF, genre) and so forth. A *between-subject design* allows us to obtain more reliable responses from users. Ultimately, in order to avoid a position-biased, the ranking of the recommendation lists was

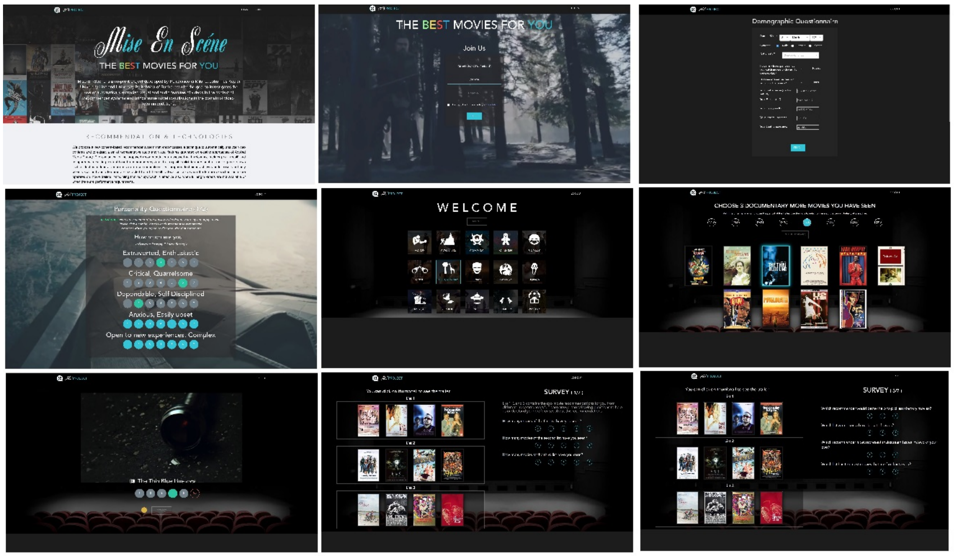


Figure 4.10: *Screenshots of the proposed MISRec web application designed for movie recommendation and running empirical studies for movies. This is an improved version of the system presented in [112].*

randomized for each user.

We used an established questionnaire used previously in a similar domain [110,182] to measure the perceived quality of recommendations across different dimensions. See Table 4.5 for further information.

4.7.3 Results

In this section, we report the qualitative results of the user-study with regards to perceived accuracy, satisfaction, personalization, diversity and novelty. Priory to the survey, the subject is invited to indicate the number of of movies in each of the recommendation list she has seen with a value from 0 to 5 (Figure 4.10 bottom middle). We only consider recommendation lists in which the user has seen at least one movie from. Therefore, if in hypothetical scenario a user chooses List 1 as the best-matching list with regard to a specific property (*e.g.,* relevance) while she has previously indicated she has seen no movie from that last, we do not consider this response valid and exclude the response from the responses. The final score for each evaluation property (perceived accuracy, satisfaction, *etc.,*) is a linear combination of the responses (scores) given by members to every question belonging to the category. The answers to questions marked with a **+** contribute positively to the final score, while scores to questions marked

Table 4.5: *List of questions proposed in [110, 182] used to measure the perceived quality of recommendations. Note that answers/scores given to questions marked with a* **+** *contribute positively to the final score, whereas scores to questions marked with a* **-** *are subtracted.*

Factor / Question (W. l. = ’Which list’, W. r. = ’Which recommender’)
Percieved Accuracy
W. l. has more movies that you find appealing? **(Q17 +)**
W. l. has more movies that might be among the best movies you see in the next year? **(Q19 +)**
W. l. has more obviously bad movie recommendations for you? **(Q6 -)**
W. r. does a better job of putting better movies on the left? **(Q9 +)**
Diversity
W. l. has more movies that are similar to each other? **(Q22 -)**
W. l. has a more varied selection of movies? **(Q7 +)**
W. l. has movies that match a wider variety of moods? **(Q13 +)**
W. l. would suit a broader set of tastes? **(Q2 +)**
Understands Me,
W. r. better understands your taste in movies? **(Q12 +)**
W. r. would you trust more to provide you with recommendations? **(Q18 +)**
W. r. seems more personalized to your movie taste? **(Q14 +)**
W. r. more represents mainstream tastes instead of your own? **(Q3 -)**
Satisfaction
W. r. would better help you find movies to watch? **(Q8 +)**
W. r. would you be more likely to recommend to your friends? **(Q16 +)**
W. l. of recommendations do you find more valuable? **(Q11 +)**
W. r. would you rather have as an app on your mobile phone? **(Q20 +)**
W. r. would better help to pick satisfactory movies? **(Q1 +)**
Novelty
W. l. has more movies you do not expect? **(Q21 +)**
W. l. has more movies that are familiar to you? **(Q4 -)**
W. l. has more pleasantly surprising movies? **(Q5 +)**
W. l. has more movies you would not have thought to consider to watch? **(Q10 +)**
W. l. provides most new suggestions? **(Q15 +)**

with a **-** contribute contrarily.

The bar plots in Figure 4.11, shows the final results along each of the under-study evaluation metrics.

Perceived Accuracy: It can be seen that tag and the SoA visual features Deep have the highest perceived accuracy with 32% and 22% of the votes which stand far ahead of the the least accurate ones, that are traditional audio and visual features BLF and AVF with 8% and 9% of the votes. The two remaining recommender Genre and SoA audio features i-vector obtained mediocre number of votes 15% and 14%. These outcomes are in agreement with the results achieved in Experimental Study A: Offline Experiment at least partially meaning that as an independent feature, the proposed SoA feature Deep show the most encouraging results compared with the rest of features (*e.g.,* compare Deep 22% with Genre 15% (approximately 50%

improvement). It was however expected that tag features as a rich semantic human-generator descriptor obtain the highest user-perceived accuracy votes as this is witnessed. The most surprising outcome *w.r.t.* perceived accuracy is the performance of Genre feature which ranked 3rd among the ranking list of recommendation performance while our offline results show the worst performance for genre.

Understands Me and Satisfaction: The votes obtained for user-perceived personalization and overall user experience with the RS are highly correlated. These two dimensions are captured by questions in Understand Me category and Satisfaction category respectively. These outcomes particularly for Satisfaction are pretty correlated with perceived accuracy and indicate that the users' notion of personalization and satisfaction are the same as accuracy and they respond these questions as belonging to the same category. Again, the results reveal a superior performance for tag and Deep features compared with the traditional audio and visual features BLF and AVF.

Diversity: The results for the perceived diversification show the best performance not by SoA visual or Audio features (as in Perceived Accuracy) rather for traditional audio features BLF with 25% of votes followed by metadata features (genre and tag) both with 20% of votes and SoA audio feature i-vector in the third place with 16% of votes. The visual features (both SoA Deep and traditional AVF) indicate the lowest perceived diversity. The results of diversity are interesting and demonstrate that audio signals are more powerful for diversification of the recommendation list in the task of video recommendation.

Novelty: The votes for novelty are slightly surprising. The traditional visual features AVF indicate the highest amount of perceived novelty gaining as much as 40% of votes, followed by tag with 29% votes. This is while genre feature which performed the highest in the offline experiment with regards to novelty, scores negative in online section *w.r.t.* perceived novelty.

Overall, we can summarize these results as follows: as a unimodal feature, the user-perceived quality of recommendation is higher with regards to the SoA visual and audio features, Deep and i-vector compared to the traditional AVF and BLF. Both of these MM feature are ranked slightly lower than semantic rich tag features as it could be excepted. In the offline experiment, we obtained the same pattern of results with the difference that the SoA audio feature i-vector outperformed tag. However, interestingly for novelty and diversity, the results in the offline experiment and the user-study do not confirm with each other perfectly; in fact it can be seen that traditional audio and visual features BLF and AVF have a promising perceived effect. The lesson that can be learned is to adoption of these established features together with other rich item descriptors (such SoA audio and visual features) and tag in the design of movie recommendation system can increase the perceived diversity and novelty of the RS. We can remind that other works previously reported contrasting results in the offline experiment and the user-study [76, 77].

4.8 Experimental Study C: Addressing the Open Research Questions

In this section, we examine the overall results obtained and provide detailed answers/viewpoints to the open research questions posed in the Introduction, Section 4.1).

4.8.1 RQ1: *video genome* vs. metadata for video recommendation

During the two above investigations, we used multimedia content-based descriptors of different nature (audio *v.s.,* visual) or (novel SoA *v.s.,* established traditional) to realize movie recommendations and contrasted the performance of such CBRS with regards a number of accuracy and beyond-accuracy measures.

In Study A, the SoA audio and visual features generally outperform metadata with regards to the three tested accuracy measures (MRR, MAP and Recall) where the difference between audio-genre or Deep-genre remained significant but not substantial *w.r.t* tag. The best cutoff value for performing a correct comparison with the user-study is @4. In offline experiment, both in the cold-start and warm-start scenarios, i-vector was the best-operating feature @4 closely followed by the Deep visual feature. For Study B on the other hand, the metadata tag feature has consistently the superior performance in Accuracy, Satisfaction and Understands me, followed by deep features. Therefore, one can note slightly contrasting results in ranking of different recommenders in terms of their performance with regards to Accuracy.

As for the result of Novelty in Study A, the metadata features provide a good performance which stands in stark contrast with the results in Study B in which with the traditional visual feature AVF exhibits the best performance. From the viewpoint of Diversity, i-vector and AVF show the best performance in offline experiment different from the user-study with BLF and metadata outperforming the rest. As stated earlier, other works previously reported similar contrasting results in the offline experiment and the user-study [76, 77].

These results provide empirical confirmation that: *The proposed state-of-the-art visual and audio features based on Deep Learning [194] and i-vectors [107] extracted from trailers can be well utilized as an alternative to conventional metadata content such as genre labels and user-generated tags to realize movie recommendation, either individually or hybridized with metadata. This is while, the traditional audio and visual features based on musical Timbre [180] (BLF) and aesthetics [135] (AVF) show a better performance in improving the beyond accuracy metrics (Novelty, Diversity) considering the results from both studies.*

4.8.2 RQ2: which audio-visual information is better

The results of both studies indicate that the SoA audio and visual features provided better recommendation quality compared with the conventional ones. Indeed, i-vector and deep features outperform BLF and AVF with regards to CS and WS Accuracy measures (see Table 4.2 and 4.3) and better user provided scores for Study B's Accuracy, Satisfaction and Understands me evaluation categories, as presented in Figure 4.11.

We observed better performance for established/traditional features AVF and BLF only with regard to some beyond accuracy metrics (diversity or novelty) in the online user study.

4.8.3 RQ3: is fusing information more effective

The outcomes of offline experiment as presented in in Tables 4.2, 4.3 and 4.4 witness that hybridization techniques based on i-vector and deep improve the unimodal results. While i-vector + deep may not always improve the scores of its individual components, i-vector + tag and deep + tag consistently rank the best in accuracy measures, with a significant improvement both over their components score and over every other feature. These big difference in Study A, lead us to believe that the proposed multimodal fusion method would be a good choice for creating better recommender systems.

4.9 Conclusions of Chapter 4

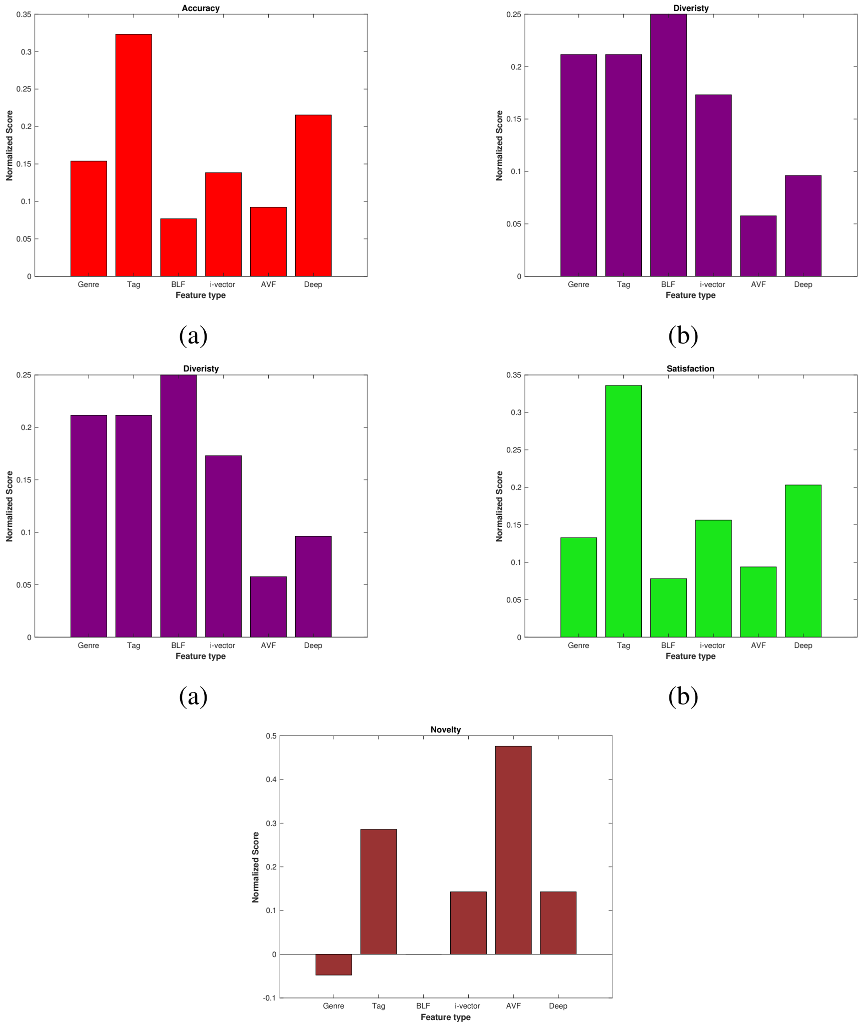
In this chapter, we presented a novel framework for multi-modal content-based movie recommendation by exploiting *rich item descriptors*. We compared them to standard metadata-based methods that use genres and tags.

The framework combines multimedia established/traditional *aesthetic visual features and audio block-level features* and novel state-of-the-art *deep visual features and i-vectors audio features*. Furthermore, we we proposed a *rank aggregation strategy* extending the Borda count approach to leverage the informativeness of different descriptors for the final improved recommendation quality. Furthermore, our system represents a practical solution to *alleviate the cold-start problem*, where metadata are unavailable or limited.

For evaluation, we performed two wide empirical studies: (i) a system-centric study to measure the offline quality of recommendations along accuracy-related (MRR, MAP, Recall) and beyond-accuracy (novelty, diversity, and coverage) performance measures, and (ii) a user-centric online experiment, measuring different subjective metrics, including relevance, satisfaction, and diversity. In both studies, we used a dataset of more than 4,000 movie trailers, which makes our approach more versatile and effective (trailers are more readily available than full movies).

The following research questions were addressed:
RQ1: *Can the exploitation of the* video genome *describing the rich item information provide better recommendations than traditional approaches that use human-generated metadata?* As our experiments have demonstrated, multimedia content can provide a good alternative to metadata with respect to accuracy and beyond accuracy measures.
RQ2: *Which of the visual and audio information play the most significant role in driving users' preference toward a video?* The most significant improvement for the accuracy metrics was achieved when using novel, state of the art approaches for audio and visual features: i-vectors and deep neural network features.
RQ3: *To which extent fusing audio and visual information can improve the quality of recommendation?* Multimodal fusion approaches have shown the best results in our studies. Audio-visual features can improve results, either when being used together or when they are combined in a hybridization scheme with human-generated metadata features.

In the work at hand, we intentionally considered only content-based recommendation models, refraining from CF approach. The main motivation is to improve our understanding of the contributions of individual and joint content descriptors to various accuracy and beyond-accuracy performance measures. As part of future work, we will look into hybridization approaches to combine the best-performing *video genome* features with CF techniques, using the proposed enhanced Borda count aggregation, but also content-boosted CF [231], factorization machines [263], and regression-based latent factor models [114]. In addition, inspired by recent psychological research in the music domain [119, 273], we plan to investigate the relationship between users' personality traits and movie preferences as well as preferences towards certain beyond-accuracy requirements like diversity or novelty of recommendations. We also plan on investigating the effect of using different labels, such as genre and tags collected from users, to build tag-aware or genre-aware deep models and i-vectors for the recommendation task.



(a)

(b)

(a)

(b)

(b)

Figure 4.11: *Real-world user study along with 5 dimension. Y-axis represents user votes in favor or against particular feature along with a certain dimension (Accuracy, Diversity etc.,). Negative values correspond to a negative opinion.*

CHAPTER 5

Evaluation and Dataset

5.1 Movie Datasets

In recent years and in presence of huge volume of video content created, shared, and consumed through various web channels and the diversity of the content, *e.g.,* user-generated videos, movies, music video clips, and so on, the role of recommender systems that automatically determine the content that a user may like has been become more prominent [16, 23, 268].

There is an obvious need for researchers and practitioners to get access to stable, large-scale, and multimodal (audio-visual) datasets of movies to test and benchmark their recommendation and personalized search and retrieval algorithms. The closest efforts to build such datasets in recent years include Netflix[1] and EachMovie datasets where both are no longer accessible. Perhaps the most valuable, still available dataset, is the MovieLens (ML) dataset [142], which contains timestamped ratings of users on movies which can be used to perform research on personalized movie search and recommendation. These preferences originate from users of MovieLens[2], a real web-based movie recommender system largely based on collaborative

[1] https://www.netflix.com

[2] http://www.movielens.org

filtering models and movie reviews. Several versions of the dataset have been released since its introduction in 2005, such as ML-100K, ML-1M, ML-10M, and ML-20M which primarily differ in terms of the size of rating, users, items and availability of user-generated tags. However, the main shortcoming of such datasets is lack of content features characterizing the audio and visual modalities of the movies. In this regard, while in multimedia signal processing and multimedia information retrieval communities, extraction of different types of audio and visual features from video content is intensely researched, the RS community for a long time has been biased with the metadata-centric notion of the “content” disregarding a wealth of information encoded in the actual signals. Although datasets like ML [142] and Yahoo! Movies WebScope dataset [5] provide different types of metadata to serve as the “content” in RS, as shown in previous chapters, these features do not fully represent the content of movies, whether be it in the form editorial metadata such as genre or user-generated tags or keywords.

Table 5.1: *Most relevant datasets created for the development of recommender systems (M - metadata, A - audio, V - video, $|\mathcal{I}|$ — number of items, $|\mathcal{U}|$ — number of users, $|\mathcal{P}|$ — number of ratings).*

Dataset	Domain	Content	Preference Type	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{P}\rvert$
MovieLens 20M [142]	movie	M	ratings [1-5]	26.7K	138.5K	20M
Yahoo! Movies WebScope [5]	movie	M	ratings [F-A+]	9K	2K	91K
LDOS-CoMoDa [192]	movie	M + context	ratings [1-5]	1K	1K	2K
Million Song [42]	music	A, M	listening events	1M (track)	1M	48M
Million Musical Tweets [143]	music	A, M	listening events	134K (track), 25K (artist)	215K	1M
LFM-1b [275]	music	M	listening events	32M (track), 3M (artist)	120K	1.1B
MMTF-14K	**movie**	**M, A, V**	**ratings [1-5]**	**13.6K**	**138.5K**	**12.4M**

Addressing specifically these limitations, the primary contribution of this works is the design and release of a large-scale publicly available multifaceted movie trailer dataset (MMTF-14K) to facilitate research on recommendation, classification and retrieval.

The results of this endeavor has been recently accepted at the software and dataset tracks of the *ACM Conference on Multimedia Systems 2018* (MMSys’18) (accepted in April 2018) [86].Part of these data is used in a MediaEval 2018 task[3](See [85] for more information)

5.2 Previous Work

In the video domain, the most widely appreciated still available dataset for recommendation tasks is the MovieLens (ML) dataset [142]. The ML

[3] http://www.multimediaeval.org/mediaeval2018/

dataset contains timestamped rating scores of users for movies. Different versions of the dataset have been released to date which differ foremost with regards to the number of ratings, users, and items: ML-100K, ML-1M, ML-10M, and ML-20M. In 2005, ML introduced tagging facilities by user enabling generation of user-generated tags/keywords in the later release of the dataset (10M and 20M). Because of the significant value that this dataset provides in investigating and validating personalization and recommendation algorithms, the ML dataset has been largely adopted by the RS community, and referenced in the research literature ever since (*e.g.,* 7,500+ references to ML in Google Scholar) [142].

Among the few other accessible video datasets, Yahoo! Movies Web-Scope dataset [5] provides a small fraction of the movie community's ratings for various movies which are on a scale from A+ to F. A large number of metadata features such as cast, crew, synopsis, genre, average ratings, awards come with the dataset, which are restricted to the movies released by November 2003 and prior. Another movie recommender dataset is the LDOS-CoMoDa dataset [192] which is a context-aware dataset providing community ratings given to movies in addition with twelve pieces of contextual information in which the movies were consumed, such as time, day type, season, weather, mood and health-condition. The dataset facilitate research on CARS however none of these movie datasets provide advanced, precomputed, audio and visual descriptors.

In music recommendation domain however, the community has been very active in producing considerably large number of openly accessible dataset containing musical features. The most well-known is the Million Song Dataset (MSD) [42], which was released in 2012 together with the MSD Challenge[4]. MSD combines various kinds of information in one million contemporary popular music pieces. Other examples are the the Million Musical Tweets Dataset (MMTD) [143] and the LFM-1b dataset [275]. Although these datasets have a different focus than our topic (movie recommendation), we consider them essential as they are related to the audio information.

We propose and release the MMTF-14K dataset which is designed to be used for building movie RS using latest advances in audio-visual content representations. The dataset is publicly available and consists of 13,623 Hollywood-type movie trailers, ranked by 138,492 users with a total of almost 12.5 million ratings. A summary of the features provided and a comparison with other datasets is presented in Table 5.1. To the best of our knowledge, the MMTF-14K dataset is the first large-scale dataset in the

[4] https://labrosa.ee.columbia.edu/millionsong/challenge

Table 5.2: *Distribution of genre labels in MMTF-14K.*

#	Genre	#pos	#neg	skewness
1	Action	1,766	11,857	6.71
2	Adventure	1,202	12,421	10.33
3	Animation	482	13,141	27.26
4	Children	592	13,031	22.01
5	Comedy	4,139	9,484	2.29
6	Crime	1,473	12,150	8.25
7	Documentary	1209	12414	10.27
8	Drama	6,592	7,031	1.07
9	Fantasy	737	12,886	17.48
10	Film-Noir	151	13,472	89.22
11	Horror	1,453	12,170	8.38
12	Musical	509	13,114	25.76
13	Mystery	754	12,869	17.07
14	Romance	2,003	11,620	5.80
15	Sci-Fi	938	12,685	13.52
16	Thriller	2,233	11,390	5.10
17	War	543	13,080	24.09
18	Western	323	13,300	41.18
Avg.		**1,505.5**	**12,118**	**18.65**

recommender systems community that provides all types of content-based descriptors in conjunction with metadata. MMTF-14K is also multifaceted allowing to develop connections with other related domains such as: *popularity prediction* — in the RS community, the common way to calculate popularity is based on how much a movie has received attention/interaction from the user community. Formally, the popularity of item i is calculated as the fraction of users who have rated item i, over the total number of users [346]. Since MMTF14K provides links to the ML ratings dataset, we can measure popularity based on number of interaction each movie has received; *genre classification* — predict movie genre by using multimedia descriptors. Given the 18 binary genre labels, the problem is in essence a multi-label classification problem; *tag-prediction (auto tagging)* — automatically predict/recommend tags given a media content. This is done by learning the association between textual multimedia features and tag keywords.

5.3 Dataset Description

5.3.1 Provided content descriptors

Apart from the from the movie trailers (which are provided via links), MMTF-14K comes with precomputed state-of-the-art features, addressing three modalities: metadata (textual), audio and visual.

Metadata descriptors

Two types of metadata descriptors are provided with MMTF-14K: (i) genre features as editorial metadata, and (ii) tag features to serve as user-generated metadata. Additionally, we provide the year of production for the movies in MMTF-14K as a side contribution. The metadata originally belong to the ML dataset which provide them in textual form (see Table 5.3). Nevertheless, in MMTF-14K these data are preprocessed and prepared as ready-to-use numerical feature vectors. The advantages of releasing metadata are multi-fold: first, they can be used in building CB or Hybrid RS as features describing items' content. For example, metadata can be used in conjunction with multimedia features using a variety of fusion or hybridization techniques; ultimately, they serve as baselines for comparing the recommendation quality with other systems.

Genre features: are represented by a 18-dimensional binary vector for each movie trailer, representing each of the 18 annotated movie categories: *Action*, *Adventure*, *Animation*, *Children*, *Comedy*, *Crime*, *Documentary*, *Drama*, *Fantasy*, *Film-Noir*, *Horror*, *Musical*, *Mystery*, *Romance*, *Sci-Fi*, *Thriller*, *War* and *Western*. The distribution of genre labels in MMTF-14K is shown in Figure 5.2. The class imbalance for each genre label is defined by the skewness ratio, given by: $Skewness = \frac{\text{negative examples}}{\text{positive examples}}$. If it is intended to use the genre labels provided by MMTF-14K for genre classification task, this classification task is in essence a *multi-label classification problem*. In such a condition, knowledge of skewness is fundamental since in multi-label classification approaches such as one-vs-rest, the class imbalance can substantially influence the performance of classifiers. In other words, for such datasets adoption of conventional classifiers and/or evaluation metrics (*e.g.,* accuracy) may not provide realistic picture of the overall classification quality [311].

Tag features: are based on a term-frequency inverse-document-frequency (tf-idf) Bag-of-Word (BoW) model. A preprocessing stage is added to the process of generating the final movie-level descriptor, involving the following operations: *(a) punctuation removal*, *(b) tokenizing and lower-case* *conversion*, *(c) word removal* for words with very high or very low frequency, *(d) stop word removal* and finally *(e) Porter stemming* [251]. After these steps, each tag feature is represented by a decimal vector of length 10,228. Note that only 9,646 of the movies were assigned tags by users. This happens in cold-start (CS) situations when a new item is added to the catalog and no metadata is assigned to it [23].

Year of Production: In addition to the above metadata, we release also the year of production for movies. They were automatically obtained by text processing of the ML dataset and extracting the years embedded in the title of movies. For a few movies, this information was not available and was added manually. The average, median and standard deviation of these values are: 1992.2, 1999 and 84.84, respectively.

Audio descriptors

Two sets of audio features are provided, representing both traditional descriptors (Block-level features) as well as current state-of-the-art (I-vectors).

Block-level features (BLFs): extracts features from audio segments of a few seconds, in contrast to frame-level features which operate on much shorter units. BLFs capture temporal aspects of an audio recording to some degree. The block-level feature framework [291] defines six features that capture: *spectral aspects* (spectral pattern, delta spectral pattern, variance delta spectral pattern), *harmonic aspects* (correlation pattern), *rhythmic aspects* (logarithmic fluctuation pattern), and *tonal aspects* (spectral contrast pattern).

I-vector features: this paradigm [83] is the current state-of-the-art representation learning technique in different audio-related domains, such as speech processing, music recommendation, and acoustic scene analysis [107, 316]. An i-vector is a fixed-length and low-dimensional representation containing rich acoustic information, which is usually extracted from short to moderate segments (usually from 10 seconds to 5 minutes) of acoustic signals. The i-vector features are computed using Mel-Frequency Cepstral Coefficients (MFCCs) frame-level features.

Visual descriptors

Similar to the audio descriptors, for the visual modality we provide both a traditional approach (Aesthetic Visual Features) and state-of-the-art descriptors that use deep neural networks (AlexNet Features).

Aesthetic Visual Features: this set of features has been proposed in [135], where they have been used for measuring the aesthetic value of coral reef pictures, while parts of these feature set are inspired from works dealing with artwork and photographic aesthetics [80, 174, 202]. This set is composed of 26 features, split into three main categories: *color based descriptors*, *texture based descriptors* and *object based descriptors*. Three early fusion schemes are presented in our collection for recommendation purposes: individual features, feature fusion according to the three main categories and full fusion with all the features. We also employ four aggregation schemes for obtaining movie level features: average, median, average + variance and median + median absolute deviation.

Table 5.3: *Characteristics of the user-rating matrix associated with MMTF-14K and ML-20M: $|\mathcal{U}|$ — number of users, $|\mathcal{I}|$ — number of items, $|\mathcal{R}|$ — number of ratings. (Note that w.r.t. Table 5.2 $|\mathcal{R}|$ = $|\mathcal{P}|$)*

dataset	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$	$\|\mathcal{R}\|$	$\frac{\|\mathcal{R}\|}{\|\mathcal{U}\|}$	$\frac{\|\mathcal{R}\|}{\|\mathcal{I}\|}$	$\frac{\|\mathcal{R}\|}{\|\mathcal{I}\|\times\|\mathcal{U}\|}$ (density)
MMTF-14K	138,492	13,623	12,471,739	90.05	915.5	0.0066
ML-20M	138,493	26,744	20,000,263	144.4	747.8	0.0054

AlexNet features: the AlexNet [194] deep neural network has been developed for scene and object recognition tasks. In our context, we use the extracted output values of the fc7 layer, as it has been shown to give good performances in a high number of tasks, some of which are related to human centered preference systems such as interestingness [98] and aesthetic ranking [186]. The same four statistical aggregation schemes as for the aesthetic visual features are employed here.

5.3.2 Dataset basic statistics

The dataset consists of 13,623 movie trailers, with an average duration of approximately 2 minutes and 22 seconds, for a total of almost 23 days of video data. 138,492 users gave 12,471,739 ratings to these trailers, thus each user rates an average of 90 videos and each video has being rated on average 915.5 times as shown in Table 5.3. As it can be seen, statistical distribution of ratings in MMTF-14K and ML-20M represented by the density of URM $density = \frac{|\mathcal{R}|}{|\mathcal{I}|\times|\mathcal{U}|}$ are similar (density = 1- sparsity).

5.3.3 Data format

The MMTF-14K dataset can be downloaded from this link[5]. The data is organized in 4 folders, one of them containing general information regarding

[5] MMTF-14K dataset: `https://mmprj.github.io/mtrm_dataset/index`.

the dataset (*Data*), while the other 3 (*Metadata descriptors*, *Audio descriptors* and *Visual descriptors*) have the preprocessed features under different aggregation schemes. All the comma-separated values (.csv) files are encoded in simple UTF-8 format, while archive files are in standard ZIP format.

Data

The *Data* folder contains two files that offer information regarding the dataset: *movie_description.csv* and *rating.csv*. The first file gives details about the trailers in this dataset, with the first column indicating the movie id, the second indicating the full title of the movie trailer while the last one representing the preferred trailer link (YouTube identifier)[6] where the trailers can be accessed. The second file presents the rating dataset after being merged with the ML-20M dataset [142][7] representing the ground truth data that can be used with this dataset for movie recommendation purposes.

Metadata descriptors

The *Metadata descriptors* folder contains three subfolders: *Genre features*, *Tag features* and *Year of Production*. All the metadata features are obtained from Movie Lens. The *Genre features* folder contains the *GenreFeatures.csv* file with every row representing a movie. The first column of this csv file represents the id of the movie trailer corresponding the the ML movie ids, while the rest of the columns represent the binary values of the genre feature vector. The *Tag features* folder has a similar structure, containing a *TagFeatures.csv* file, where movie trailers are represented as different rows. Again, the first column is the id of the movie trailer and the tag feature vector is contained in the rest of the columns.

Audio descriptors

The *Audio descriptors* folder contains two sub-folders: *Block level features* and *I-Vector features*. While the Block-level features include different fusion schemes, the I-Vector features include different parameters for the gaussian mixture model (GMM) and total variability dimension (tvDim). The *Block level features* folder has two sub-folders: *All* and *Component6*, and while the former contains the fusion scheme of all the 6 subcomponents of the BLF, the later contains each of the 6 components in separate csv files.

[6] The full link in order to access the trailers is created by: `https://www.youtube.com/watch?v=` + YouTube identifier

[7] `http://www.movielens.org`

The *All* folder contains a .csv file named *BlockLevelFeatures-All.csv*. Here, every movie trailer is represented on a different row, the first column of every row representing the id of the trailer, while the rest of the column holds the BLF feature. The *Component6* folder contains 6 .csv files, each representing a component of the BLF vector (e.g.: *BlockFeatures - Component6 - Spectral.csv*, *BlockFeatures - Component6 - SpectralContrast.csv*) with a similar structure as the previous file. The *I-Vector features* folder includes the 180 files, corresponding to the all combinations of the parameters used (e.g.: *IVectorFeatures - GMM_tvDim_fold - 16_10_1.csv*, *IVectorFeatures - GMM_tvDim_fold - 512_400_5.csv*) where the first number of the title of the files represents the number of mixture models of the GMM (16, 32, 64, 128, 256 and 512), the second the tvDim (10, 20, 40, 100, 200 and 400) and the last one the fold number (1, 2, 3, 4 and 5). Extraction of i-vectors requires building an acoustic space from the audio signals of the item in the train phase which is used to learn/extract i-vector features from each item in a subsequent stage. Since this task is dataset-dependent, in the folder Data we provide an additional folder *rating-splitted-5foldCV* which contains rating matrices for each fold and the corresponding items used to extracts i-vectors. Obviously for a comparison with other provided descriptors, the same splitting of the ratings should be used for all descriptors.

Visual descriptors

The *Visual descriptors* folder contains two subfolders: *Aesthetic features* and *AlexNet features*, each of them including different aggregation schemes for the two types of visual features. The *Aesthetic features* folder includes 4 subfolders, corresponding to the 4 aggregation schemes: *Avg* containing the average aggregation scheme, *AvgVar* with the average and variance aggregation scheme, *Med* containing the median scheme and finally *MedMad* with the median and median absolute deviation aggregation. Each of these folders contain 30 .csv files, representing the different early fusion schemes applied to these features: individual components (i.e. *AestheticFeatures - MED - Feat26Convexity*, *AestheticFeatures - AVG - Feat26Edge*), early fusion based on the 3 main types (i.e. *AestheticFeatures - MEDMAD - Type3Color.csv*, *AestheticFeatures - AVG - Type3Texture.csv*) and finally a vector containing all the component concatenated (i.e. *AestheticFeatures - MED - All.csv*). The *AlexNet features* folder has a similar structure, containing the 4 subfolders, each of them corresponding to a different aggregation scheme: *Avg, AvgVar, Med* and *MedMad*. The structure of these archives is simpler than the case for the AVF features, considering that no early fusion scheme is needed or applicable to the fc7 layer output. Therefore, only one file will be present in these folders, depending on the aggregation scheme (i.e. *AlexNetFeatures - MED - fc7*).

5.4 Ground Truth

Since the proposed MMTF-14K dataset is mainly meant for movie recommendations task but can also address a broader audience in machine learning and multimedia domain, the ground truth here is the actual rating scores provided by the user to the movies. The recommendation problem can be formulated in two ways: (i) *prediction version of problem* (*i.e.,* predict the rating value for a user-item combination) (ii) *ranking version of problem* (*i.e.,* determination of Top-k items among all items that the user would like). Both of the above approachers require rating score as ground-truth for evaluation [23].

The ground truth associated to the data was extracted from the MovieLens 20M dataset [142], also called ML v4. This released version of the dataset consists of ratings sampled throughout a large portion of the history of the ML initiative, more precisely from January 1995 to March 2015. The rating system is a "half star" system, moving away from a "whole star" only system in 2003, as a result of user demand in some surveys, thus granting users the permission to choose from 10 preference scores (0.5 to 5). Users also participated in the creation of the original tag features assigned to each movie, a function that was added to ML in December 2005. In what concerns the provided ratings, there are fewer ratings in the "half star" categories than in the "whole star" ones, most likely due to the later introduction of the 10 score system. Secondly, the distribution of ratings counts per user (*i.e.,* number of ratings given by each user to movies in the catalog) and item (*i.e.,* number of ratings given to each item by the users) is shown in Figure 5.1. As it can be seen, the pick of rating counts per user lies in the region 11-30 rating, with an average number of ratings equal to 90.05 and a maximum and minimum equal to 4,873 and 2 ratings respectively (std: 139.14). This is while for per-item rating score, majority of items have less than 10 rating scores. For example, 3,265 movies have between 1-3 ratings. The average number of ratings per item is 915.5 while the maximum and minimum equal to 63,366 and 1 rating(s) (std: 3,434.9). As it can be noted there is a large standard deviation/sharp discontinuity between two types of items, first the so called main-stream items (the items in the left side of the orange curve which attract a fair number of rating) and second the popular items (the items in the right side of orange curve which attract a large number of ratings per user). Knowledge of popularity is

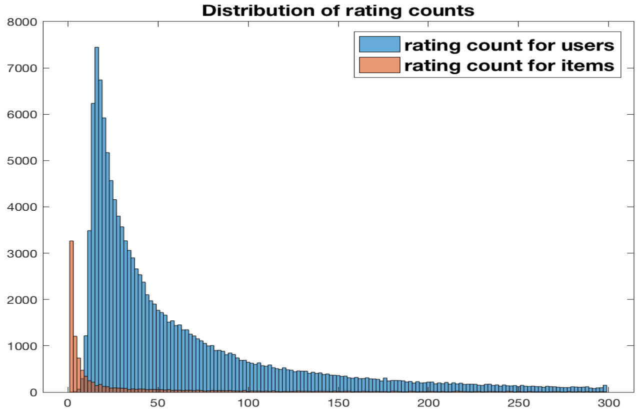


Figure 5.1: *Distribution of rating counts for users and items (x-axis: number of ratings, y-axis: count). Note that for simplicity we show the plot for rating count equal to 300.*

important because some recommendation algorithms such as collaborative filtering (CF) [23, 268] base their recommendation on the ratings obtained by community of users (instead of content descriptions as in CB) thereby promoting the chance of these popular items being more recommended. Thus, in some tasks a portion of popular items are removed from the recommendation process in order to obtain a realistic picture of the overall quality of the system [8].

Table 5.4: *Baseline performance for movie recommendation. Best results are in bold. CS: cold-start, WS: warm-start*

feature name	modality	MRR@4	MAP@4	R@4	MRR@10	MAP@10	R@10
tag (CS)	M	0.0195	0.0051	0.0042	0.0274	0.0037	0.0111
tag (WS)	M	0.0213	0.0057	0.0046	0.0294	0.0041	**0.0120**
genre	M	0.0162	0.0044	0.0039	0.0245	0.0034	0.0112
i-vec	A	**0.0233**	**0.0060**	**0.0052**	**0.0311**	**0.0042**	**0.0120**
BLF	A	0.0170	0.0045	0.0038	0.0242	0.0032	0.0097
AlexNet	V	0.0219	0.0057	0.0043	0.0296	0.0038	0.0111
AVF	V	0.0187	0.0049	0.0039	0.0263	0.0034	0.0102
i-vec + AlexNet	A + V	0.0232	0.0061	0.0051	0.0318	0.0043	0.0122
AlexNet + tag	V + M	0.0239	0.0062	0.0053	0.0325	0.0044	0.0130
i-vec + tag	A + M	**0.0266**	**0.0072**	**0.0059**	**0.0359**	**0.0049**	**0.0139**

[8] Note that in RS community, popularity is measured by the number of ratings assigned to items (*e.g.,* movies)

5.5 Baseline and Reference Results

To provide a baseline for recommender system experiments, we randomly chose a subset of 3,000 users, with the condition that each user has a minimum of 50 movie ratings in their profile, and performed a 5-fold cross validation experiment by creating 5 non-overlapping segments. *Mean Reciprocal Rank* (MRR), *Mean average precision* (MAP) and *Recall* (R) are then calculated for different cutoff values (@4 and @10) and in two different scenarios: warm-start (WS) and cold-start (CS). While the WS scenario takes into account all the tag features, the CS scenario keeps all the tag features for training the system, while on the test set only a random selection of 3% of the tag features are kept. The cold-start scenario is supposed to simulate real-world conditions, by acknowledging the fact that some movies, especially the newer or less popular ones, have a small set of user input data, therefore have fewer tags attached to them. The corresponding code for calculating the MRR, MAP and R values is available with the dataset.

The results for each default extracted descriptor are presented in Table 5.4, along with the results for the best performing late fusion combinations of these features. The fusion scheme is based on the Bodra count method [34] to fuse ranking results of different recommenders into a unified ranking of videos. As it can be seen the SoA i-vec (audio) and Deep AlexNet (visual) have supervisor performance compared with traditional BLF (audio) and AVF (visual) descriptors. It can be also noted that while both SoA audio and visual descriptors have a higher quality *w.r.t* the genre recommender, the i-vec has also a superior performance compared to semantic-rich tag (in both CS WS) *w.r.t* all evaluation metrics and across all cut-off values.

5.6 Evaluating video and music recommender systems

The evaluation of system's effectiveness is a core issue in RS. In order to decide about the most suitable algorithm for a particular task/domain, an application designer would require to select a set of properties that impact the success of the RS in the context of a specific application [134]. The identification of the these properties is inherently linked with the goal of a RS, which is by nature two-fold: (1) *user-centric goal:* It means the RS is able to suggest user items that she is interested in but she did not know about or did not know how to ask for (2) *business-centric goal:* to promote *product sale* and increase *revenue*.

Although the main focous of this thesis is on video recommender systems, many of the presented evaluation metrics are used for evaluating generic recommender systems and some are specific to music and videos *e.g.,* sequence-aware evaluations metrics.

In general, the topic of evaluation in RS has its roots in a few neighboring fields, such as machine learning (cf. rating prediction) and information retrieval (cf. "retrieving" items based on implicit "queries" given by user preferences). Novel measures that are tailored to the recommendation problem have emerged in recent years. We can classify the evaluation metrics in RS in two broad categories of (i) accuracy metrics and, (ii) beyond-accuracy metrics. In the following we review each of these metrics and review relative merits of each metric.

5.6.1 Accuracy metrics

Predictive accuracy metrics (aka error metrics)

These metrics measure *how correct* or *incorrect* the RS is in predicting what the user would have rated an item. Practically speaking these metrics act in similar manner to leave-one-out cross validation for dataset partitioning and evaluation (See [122] Ch. 6) in which the rating of one user-item pair is covered up and we evaluate how the RS is good in predicting it.

Mean absolute error (MAE) is the simplest metric for evaluating the prediction power of recommender algorithms. It computes the average absolute divergence of the predicted ratings from the actual ratings provided by users [151] defined as follows

$$MAE = \frac{1}{|T|} \sum_{r_{u,i} \in T} |r_{u,i} - \hat{r}_{u,i}| \tag{5.1}$$

where $r_{u,i}$ and $\hat{r}_{u,i}$ respectively denote the actual and the predicted ratings of item i for user u. MAE sums over the absolute prediction errors for all ratings in a test set T.

Root mean square error (RMSE) is another common error metric that is computed as:

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{r_{u,i} \in T} (r_{u,i} - \hat{r}_{u,i})^2} \tag{5.2}$$

Lower MAE and RMSE correspond to higher prediction accuracy. Since the error term in RMSE is squared before summation, it tends to penalize large errors more than smaller ones. This can be for example useful to catch huge occasional errors by an algorithm. The main merits of error metrics is on their simplicity and that many public datasets have published the results of algorithms *w.r.t.* MAE and RMSE thereby facilitating comparative evaluation. The main shortcoming of these metrics is that the error can be dominated by irrelevant parts of the product space. In other words, the errors metrics treat all rating scores equally with disregard to their positions in the recommendation list where in practice the main goal of the RS is to find a small number of item that are likely to be liked by a given user [215]. For this reason, these metrics have received huge criticism in the community of RS in recent years.

Decision-support metrics (aka classification accuracy metrics)

Suppose our movie RS is using a 5-scale ratings system ([0-5]) and for a movie with actual rating of 2.5, an algorithm has predicted it as 4 while another one predicted it as 0.5. Even though the prediction error in the first case is lower implying a better predictive accuracy, the cost of the wrong prediction in the first case is higher since the user have wasted her time watching a movie she does not like whereas in the second case she would have not watched it anyway. The name decision-support metrics is given to this class of metrics since they help users in their decision-making to identify good items from bad items. Precision and recall are the most popular metrics based on this property.

Precision at top K recommendations (P@K): Precision measures the percentage of recommended items that are relevant. $P_u@K$ for each user u is computed as

$$P_u@K = \frac{|L_u \cap \hat{L}_u|}{|\hat{L}_u|} \tag{5.3}$$

where L_u denotes the set of relevant items for user u in T and $\hat{L}_u$ is the recommendation list of size K for user u in T, T being the test set.

Recall at top K recommendations (R@K): Recall measures the percentage of relevant items that are selected. For a user u, $R_u@K$ is defined as:

$$R_u@K = \frac{|L_u \cap \hat{L}_u|}{|L_u|} \tag{5.4}$$

Averaging the individual precision and recall over all users in the test set, we obtain the mean precision and recall.

Intuitively, precision cares about return useful items so that the user's time is not wasted. Recall is about not missing useful items. The assumption behind recall is that the user will have time to go through the recommendation list and filter bad items out. While precision *typically* decreases with increase of K, recall always grows with K. There are different ways to combine them using the F_1 measure and R-Precision among others [22, 215].

Rank-aware Top-K metrics

Rank-aware top-K metrics in RS focus on the the relative ordering/position of the items within the top-K list, in other words they care about where in the recommendation list (high or low) the relevant items are. For example, a recommender algorithm that takes optimizing such property into consideration, would differentiate between the fact that if a user likes the movie "the Shindler's list" more than "Star wars", where both movies are user's favorites and therefore places the former higher in the recommendation list. For a fixed K size, the decision-support metrics cannot capture this desired property.

Reciprocal rank (RR) is one of the simplest rank-aware metrics that measures where in recommendation list, the ***first*** relevant item is. It is defined as following

$$MRR_u = \frac{1}{rank_u} \tag{5.5}$$

where $rank_u$ is the position of the first relevant recommended item for user u. If no relevant items are recommended, the value will be 0.

Average precision at top K recommendations (AP@K) is a popular rank-based metric measure by computing the arithmetic mean of the precision values obtained at the positions corresponding to relevant documents/items. $AP@K$ is formally defined as

$$AP_u@K = \frac{1}{N} \sum_{i=1}^{K} P@i \cdot rel(i) \tag{5.6}$$

where $rel(i)$ is an indicator signaling if the i^{th} recommended item is relevant, ***i.e.,*** $rel(i) = 1$, or not, ***i.e.,*** $rel(i) = 0$; N is the total number of relevant items. Note that AP implicitly incorporates recall, because it also considers the relevant items not in the recommendation list.

While the above definition of $AP@K$ is widely employed in the evaluation of information retrieval systems, in the recommender systems community, another variation of $AP@K$ has gained popularity recently in the context of recommendation challenges organized by Kaggle [2] followed by several other research works, consider for example [27, 227]. This variation of average precision for the top K recommendations ($AP@K$) is given by

$$AP_u@K = \frac{1}{\min(K,N)} \sum_{i=1}^{K} P@i \cdot rel(k) \tag{5.7}$$

in which N is the total number of relevant items and K is the size of recommendation list. The motivation behind the minimization term is to prevent the AP scores to be unfairly suppressed when the size of recommendation list is not high enough to capture all the relevant items.
Discounted cumulative gain (DCG) is a formally defined as

$$DCG_u = \sum_{k=1}^{K} \frac{rel_k}{disc(k)} \tag{5.8}$$

where rel_k is the utility of item k typically set to a non-negative function of the relevance such as non-negative ratings, $rel_k = r_{uk}$ (r_{uk} is the true rating) and $disc(k)$ is the discount term. The standard choice for the discount is logarithmic discounting by setting $disc(k) = log_2(k+1)$. The rationale behind this is to penalize relevant items that appear at a lower ranked position in the recommendation list, since highly relevant items are more useful for the user. An alternative formulation of DCG used by some web search companies, Kaggle and in RS articles is the following [2, 335]:

$$DCG_u = \sum_{k=1}^{K} \frac{2^{rel_k} - 1}{disc(k)} \tag{5.9}$$

The *Normalized Discounted cumulative gain (NDCG)* normalizes DCG by the ideal DCG (IDCG), which is simply the DCG measure of the best ranking result [22, 327]

$$NDCG_u = \frac{DCG_u}{IDCG_u} \tag{5.10}$$

The best achievable NDCG is 1. Averaging RR and AP and NDCG over the entire set of users in the test T gives mean reciprocal rank (MRR) and mean average precision (MAP) and overall NDCG.

5.6.2 Beyond-Accuracy metrics

While adopting a RS algorithm besides accuracy, other properties of the selected items should be taken into consideration since recommendation accuracy alone is not sufficient; the evaluation metrics in this category are meant to satisfy the secondary goals of RS and are as following:

Novelty: Novelty is a highly desirable feature for a RS from both business and user perspective. As for the business perspective as mentioned in Section 2.2.2 the future of business is by exploiting the niche market (less popular products) [29]. From the user perspective, novelty is also an important aspect since the purpose of recommendation is inherently linked to the notion of *discovery*. Recommendation does not make much sense when it exposes user to a relevant experience that she would have found or thought by herself anyway;

To measure novelty, commonly the self-information of the recommended items relative to their global popularity is used

$$novelty = \frac{1}{|U|}\sum_{u \in U}\sum_{i \in L_u}\frac{-\log_2 pop_i}{N} \tag{5.11}$$

where pop_i is the popularity of item i measured as percentage of users who rated i [349, 353].

Diversity: RS typically suggest a user with a list containing K items. If all the item in the suggestion list are highly similar, it risk that a dissatisfied user in one, may also dislike all other items. When the recommendation list contains diverse item types, it increase the chance of the user liking at least one of the item. diversity can be defined in several ways. One of the most common is to compute pairwise distance between all items in the recommendation set, either averaged [354] or summed [300]. In the former case, the diversity of a recommendation list K is calculated as follows:

$$diversity(K) = \frac{\sum_{i \in K}\sum_{j \in K \setminus i} dist_{i,j}}{|K| \cdot (|K| - 1)} \tag{5.12}$$

where $dist_{i,j}$ is the some distance measures between items i and j, for example inverse cosine similarity [267], inverse Pearson correlation [318], or Hamming distance [175].

5.6.3 Sequence-Aware multimedia evaluation metrics

This category of evaluation metrics have the goal of assessing the capability of the recommender in providing proper transitions between subsequent songs or videos for which the conventional error or accuracy metrics may not be able to capture transition property. There is hence a need for *sequence-aware evaluation* measures [278]. Other metrics such as *average log-likelihood* have been proposed to better model the transitions [68, 69]. In this regard, when the goal is to suggest a sequence of items, alternative *multi-metric* evaluation approaches are required to take into consideration multiple quality factors. Such evaluation metrics can consider the ranking order of the recommendations or the internal coherence or diversity of the recommended list as a whole. In many scenarios, adoption of such quality metrics can lead to a trade-off with accuracy which should be balanced by the RS algorithm [255].

5.7 Conclusions of Chapter 5

In this chapter, we released the MMTF-14K dataset, a dataset consisting of 13,623 Hollywood-like movie trailers which are rated by more than 138,492 users. The primary scope of this dataset is to support the development of movie recommender systems, and to the best of our knowledge, this is the first large-scale dataset in the recommender systems community that provides all types of content-based descriptors in conjunction with metadata. However, these data go beyond the recommending scenario thanks to its rich content. It can also be used for tasks such as popularity prediction, tag prediction and genre classification. Apart from the data, we are also releasing some baseline results to allow further benchmarking. The data is publicly available[9]. In addition, we presented a number of evaluation metrics that are commonly used for evaluating video recommender systems.

In addition, we presented the main evaluation metrics that are used to validate the quality of video recommendation models by taking into consideration various user-centric and business-centric perspectives.

[9] MMTF-14K dataset: `https://mmprj.github.io/mtrm_dataset/index`

CHAPTER 6

Conclusion

6.1 Summary of Thesis

In this PhD thesis, approaches to automated extraction of information related to videos from the multimedia content have been elaborated and techniques to leverage this information in recommender systems have been analyzed and discussed. The presented approaches address the problems of improving the quality of video recommendation in cold-start and warm-start situations, improving the quality of recommendation in presence of rich source of information such as the one based on CF or CBF models using metadata, assessing the perceived utility of movies recommendation when recommendations are powered by multimedia alone or in conjunction with metadata among others. In this chapter, we first summarize and conclude the work conducted in this thesis. Afterward, we present the future research directions for extending this work.

6.2 Main Contributions

Over the past three and a half years, my primary goal has been uniting the fields of recommender system and multimedia information retrieval/signal processing. Although, the focus of this thesis is on video item but the ideas presented can be effortlessly transfered to many neighboring domains (*e.g.,* images, music, Web) as well areas where the target product is not really a media content (*e.g.,* social networks and tourism). A discussion on the main contributions of this work with respect the research goals established in Chapter 1 is presented in the following:

6.2.1 Characterizing recommender systems exploiting multimedia content

In Chapter 2, we have presented a novel and in-depth survey of the state of the art in recommender systems exploiting multimedia content.

- In section 2.1, we have first presented foundations of multimedia processing and recommender systems and explained the basic concepts in each of the two fields. We elaborated and discussed different models of feature extraction for each of the audio and visual signals in addition to models to represent features, aggregate them over time and fuse them together. We also presented different types of CBF and CBF+CF recommendation algorithms that can incorporate a wealth of information encoded in the video signals. Some of these models include pure CBF based on K-nearest neighbor (KNN), attribute-to-feature mapping (AFM), regression-based latent-factor model (RLFM) and factorization machines (FM).

- In section 2.1.2, we have presented a general framework for building a content-based RS exploiting the multimedia content and described basic practical details of different processing stages involved to extract useful information from each modality.

- In section 2.2, we have presented a novel and exhaustive review of the state of the art in recommender systems exploiting multimedia content and presented a categorization of research works exploiting multimedia content with respect to the objectives and challenges they address.

6.2.2 Video recommendation by exploiting the visual content

Based on a number of my publications [87,88,90,91,94,112], in chapter 3, we have presented a new kind of CBF technique that filters videos according to their visual characteristics defined by the amount of color variation, motion and lighting key where these features correspond with the stylistic elements used in applied media aesthetics [344] used to convey communication effects and to simulate different feelings in the viewers. I have presented, analyzed and elaborated on solving several video recommendation problems by exploiting low-level visual features extracted from the visual channel of movies. Several subproblems are addressed in this context including analyzing the quality of different low-level visual features in offline experiments in both cold-start and warm-start scenarios, investigating the perceived utility of recommendation using low-level visual features alone or hybridized with high-level semantic features, analyze the offline quality of recommendation under hybridization techniques based on canonical correlation analysis (CCA) and factorization machines (FM). The results indicate that low-level (stylistic) visual features (*e.g.,* colors, light, motion) can be well and even more representative of some of high-level semantic features (*e.g.,* genre) in providing content-based recommendations.

6.2.3 Movie trailer v.s. full movies

Movies and the corresponding trailers are not produced necessarily by the same director. For movies, the director applies various filmmaking conventions ("mise-en-scene" elements) to influence the believability of a film in the eyes of a viewer. However, while movies are made with a natural pace to increase the believability of the scenes, trailers (aka previews) are used as an advertisement tool for a feature film that is planned to be exhibited on TV or at the cinema. As such, trailers can be made with a lot of abrupt cutting and merging of the scenes since their production goal is mainly to "excite the users" to watch the full movies. The correspondence of the trailers with their full-movies can be dependent on the genre, country of their production and other factors.

In section 3.2, we built a RS in order to measure the "usefulness" of movie recommendation *w.r.t* the content extracted from the much shorter version (*i.e.,* trailers). Our goal was to measure if the same quality of recommendation can be obtained based on multimedia content extracted from each of the two. Based on a number of visual feature identified in the community of multimedia processing, the results of our research indicate that low-level features extracted from movie trailers are well correlated with the corresponding features extracted from full-length movies. Comparable quality of recommendation are obtained when a CBRS is powered by contents extracted from each of the two modalities. These results are interesting (although unexpected) but if they are also confirmed for larger variety of features including auditory features it can be very useful for practical movie recommendation applications (like those developed in companies) since full-movies are not easily accessible everywhere and processing of full-length movies is computationally much more demanding than the much shorter version trailers. On advancing this research line, I have proposed an extension of this research line as a research challenge in MediaEval Benchmarking Initiative for Multimedia Evaluation 2018 (task name: Recommending Movies Using Content: Which content is key?) [1].

6.2.4 Video recommendation by exploiting the latest advances in audio and visual content analysis, both in offline experiments and via user-studies in a real deployed system

Based on my currently under-review paper at journal of User Modeling and User-Adapted Interaction (UMUAI), I have deployed a CB movie RS that leverages the latest state-of-the-art visual and audio features along with metadata, in order to build *rich item descriptions*. We refer to this rich content information as the *Video Genome* (similar to a biological DNA), since it can be considered as the footprint of both content and style of a video [52]. I have further proposed an improved version of Borda count method to hybridize the ranking output of different recommenders and show that under this hybridization, the quality of recommendation is significantly improved compared when the standard form of Borda count method or when the recommenders are used individually. An extensive set of experiments are carried out by conducting two large-scale empirical studies: (i) a system-centric study to measure the offline quality of recommendations along with various accuracy-related metrics such as MRR, MAP, Recall and beyond-accuracy novelty, diversity, and coverage, and (ii) a user-centric experiment, measuring different subjective metrics, including relevance, satisfaction, and diversity.

The results of the experiments are indicative of the fact that multimedia features can provide a good alternative to metadata with regards to both accuracy measures and beyond accuracy measures. The most significant improvement for the accuracy metrics was achieved when using novel, state of the art approaches for audio and visual features: i-vectors and deep neural network features. Multi-modal fusion under the proposed Borda count method shows the best results in our studies.

[1] http://www.multimediaeval.org/mediaeval2018/

6.2.5 Present a novel large-scale dataset to facilitate research on movie recommendation and retrieval

Based on my publication [86] (accepted in April 2018), I designed and released a large-scale multifaceted movie trailer dataset (MMTF-14K) to facilitate research on recommendation, classification and retrieval. MMTF-14K contains audio and visual descriptors in addition to ratings and metadata for 13,623 Hollywood-type movie trailers. Compared with existing datasets, this dataset is first large-scale and stable dataset in the community of RS which provides all types of precomputed content-based descriptors in conjunction with metadata in numerical feature format. Part of these data is used in a MediaEval 2018 task[2].

6.3 Outlook and Future Works

The techniques underlying the current implementation of Video Genome presented in Chapter 4, which was elaborated in the context of this PhD thesis as the latest version of our movie recommender system powered by multimedia content, although giving promising results for different information sources, still leave room for improvement in various directions. As for the task of improving recommendation quality, pursuing strategies to harvest multimedia-related information obtained from the video content, from exploring different audio and visual features such as the ones based on recent deep network models to state of the art approaches for building video-level descriptors and hybridization approaches to combine the heterogeneous multimedia content features would probably the steps that result in recommendation qualities superior to the one presented by Video Genome. Thus, building different *video-specific descriptors* is one of the next steps that should be taken. How to leverage these video-specific descriptors algorithmically in a content-driven RS powered by CBF or CBF+CF models is the next immediate related steps that should be studied.

Another line of research would be to further explore is to investigate the relation between video content and emotion. In general, I believe that movies have potentially the power to engage users psychologically and cognitively with the story to a high extent, different *e.g.,* from music which have a much shorter duration and are emotionally/cognitively often less engaging [127]. In [160], the authors witnessed this potential by evaluating the effect of positive affect in patients, inducted by ten-minute comedy films. In [331], the authors showed that compared to other media types

[2] http://www.multimediaeval.org/mediaeval2018/

like music, movies are one of the best methods to elicit emotions. In this line, pursuing an strategy to track the emotion of users during the movie-watching process and identifying the key emotional peaks in the movies or the way the emotions are changed in the course of a movie and which emotion are more evoked, can be the key to success of the field. Previous work on music domain, has proven the direct relation between musical content and emotions. I believe the same holds true for movies with respect to the evoked emotions. In addition, being able to replace the holistic preference of users given to entire movies with a "compositive" one specified to different sections in the movie can be a key to the advancement of this field. I believe a variety of wearable devices or sensors can be used to obtain this kind of preferences from users *e.g.,* the signals obtained from brain electroencephalogram (EEG) or information obtained from user eye/gaze tracking and recognition of facial emotion can be interesting steps that can taken in this direction. Doing so would presumably yield more accurate recommendation.

On the application side, there is a potential to leverage the result of this thesis for building video recommendation systems for kids or an interactive system allowing the user to actively show her interest movies to the system which I studied in the works [95, 96].

Bibliography

[1] Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020 white paper. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. Accessed: 2016-12-1.

[2] Kaggle Official Homepage. [last accessed March 11, 2018].

[3] Speaker verification: Text-dependent vs. text-independent. https://www.microsoft.com/en-us/research/project/speaker-verification-text-dependent-vs-text-independent/. Accessed: 2018-3-8.

[4] Tubularinsights: 500 hours of video uploaded to youtube every minute [forecast]. http://tubularinsights.com/hours-minute-uploaded-youtube/. Accessed: 2018-19-1.

[5] Yahoo!: Webscope movie data set (version 1.0). http://research.yahoo.com/. Accessed: 2018-03-01.

[6] Spatial resolution, tutorial points. https://www.tutorialspoint.com/dip/spatial_resolution.htm, 2017. Accessed: 2017-10-17.

[7] Bilvideo-7 software: Mpeg-7 visual descriptor. http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo-7/Software.htmle, 2018. Accessed: 2018-01-2.

[8] Iva software: Mpeg-7 visual descriptor. http://image.ntua.gr/iva/tools/vde, 2018. Accessed: 2018-01-2.

[9] Mpeg-7 visual descriptor official homepage. https://mpeg.chiariglione.org/standards/mpeg-7/visual, 2018. Accessed: 2018-01-2.

[10] Vlfeat open source library of popular computer vision algorithms. http://www.vlfeat.org/api/fisher-fundamentals.html, 2018. Accessed: 2018-01-4.

[11] Fabian Abel, András Benczúr, Daniel Kohlsdorf, Martha Larson, and Róbert Pálovics. Recsys challenge 2016: Job recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 425–426. ACM, 2016.

Bibliography

[12] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. Recsys challenge 2017: Offline and online evaluation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 372–373. ACM, 2017.

[13] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer, 1999.

[14] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[15] Panagiotis Adamopoulos and Alexander Tuzhilin. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):54, 2015.

[16] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[17] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[18] Deepak Agarwal and Bee-Chung Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 19–28. ACM, 2009.

[19] Charu C Aggarwal. Content-based recommender systems. In *Recommender Systems*, pages 139–166. Springer, 2016.

[20] Charu C Aggarwal. Context-sensitive recommender systems. In *Recommender Systems*, pages 255–281. Springer, 2016.

[21] Charu C Aggarwal. Ensemble-based and hybrid recommender systems. In *Recommender Systems*, pages 199–224. Springer, 2016.

[22] Charu C Aggarwal. Evaluating recommender systems. In *Recommender Systems*, pages 225–254. Springer, 2016.

[23] Charu C Aggarwal. An introduction to recommender systems. In *Recommender Systems*, pages 1–28. Springer, 2016.

[24] Charu C Aggarwal. Model-based collaborative filtering. In *Recommender Systems*, pages 71–138. Springer, 2016.

[25] Charu C Aggarwal. Neighborhood-based collaborative filtering. In *Recommender Systems*, pages 29–70. Springer, 2016.

[26] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[27] Fabio Aiolli. Efficient top-n recommendation for very large scale binary rated datasets. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 273–280. ACM, 2013.

[28] Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Elisa Ricci, and Nicu Sebe. Viraliency: Pooling local virality. 2017.

[29] Chris Anderson. *The long tail: Why the future of business is selling less of more*. Hachette Books, 2006.

[30] Evlampios Apostolidis and Vasileios Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6583–6587. IEEE, 2014.

[31] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[32] Mohamad Hasan Bahari, Mitchell McLaren, David A van Leeuwen, et al. Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, 34:99–108, 2014.

[33] Mohamad Hasan Bahari, Rahim Saeidi, David Van Leeuwen, et al. Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7344–7348. IEEE, 2013.

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

[35] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895–904. ACM, 2008.

[36] Ilaria Bartolini, Vincenzo Moscato, Ruggero G Pensa, Antonio Penta, Antonio Picariello, Carlo Sansone, and Maria Luisa Sapino. Recommending multimedia objects in cultural heritage applications. In *International Conference on Image Analysis and Processing*, pages 257–267. Springer, 2013.

[37] Christine Bauer and Alexander Novotny. A consolidated view of context for intelligent systems. *Journal of Ambient Intelligence and Smart Environments*, 9(4):377–393, 2017.

[38] Juan Pablo Bello. Measuring structural similarity in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2013–2025, September 2011.

[39] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5):1035–1047, 2005.

[40] Hal Berghel. Cyberspace 2000: Dealing with information overload. *Communications of the ACM*, 40(2):19–24, 1997.

[41] Thierry Bertin-Mahieux, Douglas Eck, François Maillet, and Paul Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

[42] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, volume 2, page 10, 2011.

[43] Chidansh Amitkumar Bhatt and Mohan S Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011.

[44] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. Semantic Audio Content-based Music Recommendation and Visualization Based on User Preference Examples. *Information Processing & Management*, 49(1):13–33, 2013.

[45] Dmitry Bogdanov, Joan Serrà, Nicolas Wack, Perfecto Herrera, and Xavier Serra. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4):687–701, Aug 2011.

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

[47] David Bordwell, Kristin Thompson, and Jeff Smith. *Film art: An introduction*, volume 7. McGraw-Hill New York, 1997.

[48] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

[49] Matthias Braunhofer, Marius Kaminskas, and Francesco Ricci. Location-aware music recommendation. *International Journal of Multimedia Information Retrieval*, 2(1):31–44, 2013.

[50] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[51] Thomas M Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. High-performance ocr for printed english and fraktur using lstm networks. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 683–687. IEEE, 2013.

[52] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. The video genome. *arXiv preprint arXiv:1003.5320*, 2010.

[53] Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 391–400. ACM, 2010.

[54] Warren Buckland. What does the statistical style analysis of film involve? a review of moving into pictures. more on film history, style, and analysis. *Literary and Linguistic Computing*, 23(2):219–230, 2007.

[55] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[56] Georg Buscher and Andreas Dengel. Gaze-based filtering of relevant document segments.

[57] Georg Buscher, Andreas Dengel, and Ludger Van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–394. ACM, 2008.

[58] Iván Cantador and Paolo Cremonesi. Tutorial on cross-domain recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys'14, pages 401–402, New York, NY, USA, 2014. ACM.

[59] Iván Cantador, Martin Szomszor, Harith Alani, Miriam Fernández Sánchez, and Pablo Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *CEUR Workshop Proceedings*. Yannis Avrithis, 2008.

[60] Oscar Celma. Music recommendation. In *Music Recommendation and Discovery*, pages 43–85. Springer, 2010.

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

[62] Òscar Celma and Pedro Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, page 5. ACM, 2008.

[63] Òscar Celma and Perfecto Herrera. A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 179–186, New York, NY, USA, 2008. ACM.

[64] Shih-Fu Chang, Thomas Sikora, and Atul Purl. Overview of the mpeg-7 standard. *IEEE Transactions on circuits and systems for video technology*, 11(6):688–695, 2001.

[65] Chih-Ming Chen, Ming-Feng Tsai, Jen-Yu Liu, and Yi-Hsuan Yang. Using emotional context from article for contextual music recommendation. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 649–652. ACM, 2013.

[66] Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang, Ming-Chun Tien, Winston H. Hsu, and Ja-Ling Wu. Sheepdog: Group and tag recommendation for flickr photos by automatic search-based learning. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 737–740, New York, NY, USA, 2008. ACM.

[67] Li Chen and Pearl Pu. Survey of preference elicitation methods. Technical report, 2004.

[68] Shuo Chen, Josh L Moore, Douglas Turnbull, and Thorsten Joachims. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 714–722. ACM, 2012.

[69] Shuo Chen, Jiexun Xu, and Thorsten Joachims. Multi-space probabilistic sequence modeling. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 865–873. ACM, 2013.

[70] Yuqiang Chen, Ou Jin, Gui-Rong Xue, Jia Chen, and Qiang Yang. Visual contextual advertising: Bringing textual advertisements to images. In *AAAI*, 2010.

[71] An-Jung Cheng, Yan-Ying Chen, Yen-Ta Huang, Winston H Hsu, and Hong-Yuan Mark Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 83–92. ACM, 2011.

[72] Zhiyong Cheng and Jialie Shen. Venuemusic: a venue-aware music recommender system. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1029–1030. ACM, 2015.

[73] Zhiyong Cheng and Jialie Shen. On effective location-aware music recommendation. *ACM Transactions on Information Systems (TOIS)*, 34(2):13, 2016.

[74] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. *arXiv preprint arXiv:1611.05358*, 2016.

[75] Mihai Gabriel Constantin and Bogdan Ionescu. Content description for predicting image interestingness. In *Signals, Circuits and Systems (ISSCS), 2017 International Symposium on*, pages 1–4. IEEE, 2017.

[76] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):11, 2012.

[77] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. User-centric vs. system-centric evaluation of recommender systems. In *IFIP Conference on Human-Computer Interaction*, pages 334–351. Springer, 2013.

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

[79] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[80] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, 2006.

[81] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[82] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

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

[84] Yashar Deldjoo, Eghbal-zadeh Hamid Constantin, Mihai Gabriel, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018.

[85] Yashar Deldjoo, Mihai Gabriel Constantin, Thanasis Dritsas, Markus Schedl, and Bogdan Ionescu. The mediaeval 2018 movie recommendation task: Recommending movies using content. In *MediaEval 2018 Workshop*, 2018.

[86] Yashar Deldjoo, Mihai Gabriel Constantin, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. Mmtf-14k: A multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018.

[87] Yashar Deldjoo, Paolo Cremonesi, Markus Schedl, and Massimo Quadrana. The effect of different video summarization models on the quality of video recommendation based on low-level visual features. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, page 20. ACM, 2017.

[88] Yashar Deldjoo, Mehdi Elahi, and Paolo Cremonesi. Using visual features and latent factors for movie recommendation. CEUR-WS, 2016.

[89] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, and Pietro 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.

[90] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, 5(2):99–113, 2016.

[91] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Farshad Bakhshandegan Moghaddam, and Andrea Luigi Edoardo Caielli. How to combine visual features with tags to improve movie recommendation accuracy? In *International Conference on Electronic Commerce and Web Technologies*, pages 34–45. Springer, 2016.

[92] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. Using visual features based on mpeg-7 and deep learning for movie recommendation. *International Journal of Multimedia Information Retrieval*, pages 1–13.

[93] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. Toward building a content-based video recommendation system based on low-level features. In *International Conference on Electronic Commerce and Web Technologies*, pages 45–56. Springer, 2015.

[94] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, Paolo Cremonesi, and Franca Garzotto. Toward effective movie recommendations based on mise-en-scène film styles. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*, pages 162–165. ACM, 2015.

[95] Yashar Deldjoo, Cristina Fra, Massimo Valla, and Paolo Cremonesi. Letting users assist what to watch: An interactive query-by-example movie recommendation system. 2017.

[96] Yashar Deldjoo, Cristina Fra, Massimo Valla, Antonio Paladini, Davide Anghileri, Mustafa Anil Tuncil, Franca Garzotta, and Paolo Cremonesi. Enhancing children's experience with recommendation systems. In *Workshop on Children and Recommender Systems (KidRec'17)-11th ACM Conference of Recommender Systems*, 2017.

[97] Jordan E DeLong, Kaitlin L Brunick, and James E Cutting. Film through the human visual system: finding patterns and limits. *Social science of cinema. New York, NY: Oxford University Press. http://people. psych. cornell. edu/jec7/pubs/socialsciencecinema. pdf*, 2012.

[98] Claire-Hélène Demarty, Mats Viktor Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc QK Duong, Frédéric Lefebvre, et al. Mediaeval 2016 predicting media interestingness task. In *MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[99] James J Deng, Clement HC Leung, Alfredo Milani, and Li Chen. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):4, 2015.

[100] Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.

[101] Diana Deutsch. *Psychology of music*. Elsevier, 2013.

[102] Anind K Dey and Gregory D Abowd. *Providing architectural support for building context-aware applications*. PhD thesis, College of Computing, Georgia Institute of Technology, 2000.

[103] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[104] Justin Donaldson. A hybrid social-acoustic recommendation system for popular music. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 187–190, New York, NY, USA, 2007. ACM.

[105] Chitra Dorai and Svetha Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE multimedia*, 10(2):15–17, 2003.

[106] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup'11. In *Proceedings of the 2011 International Conference on KDD Cup 2011-Volume 18*, pages 3–18. JMLR. org, 2011.

[107] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer. CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep cnns. Technical report, DCASE2016 Challenge, 2016.

[108] Hamid Eghbal-zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. I-Vectors for timbre-based music similarity and music artist classification. In *Proceedings of the 16^{th} International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015.

[109] Hamid Eghbal-Zadeh, Markus Schedl, and Gerhard Widmer. Timbral modeling for music artist recognition using i-vectors. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1286–1290. IEEE, 2015.

[110] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM.

[111] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2011.

[112] Mehdi Elahi, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Leonardo Cella, Stefano Cereda, and Paolo Cremonesi. Exploring the semantic gap for movie recommendations. In *Proceedings of the Eleventh ACM conf. on Recommender Systems*, pages 326–330. ACM, 2017.

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

[114] Asmaa Elbadrawy and George Karypis. User-specific feature-based similarity models for top-n recommendation of new items. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):33, 2015.

[115] Daniel PW Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, volume 7, pages 339–340, 2007.

[116] Alexander Felfernig, Michael Jeran, Gerald Ninaus, Florian Reinfrank, and Stefan Reiterer. Toward the next generation of recommender systems: applications and research challenges. In *Multimedia services in intelligent environments*, pages 81–98. Springer, 2013.

[117] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Popular music retrieval by detecting mood. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 375–376. ACM, 2003.

[118] Ignacio Fernandez Tobias, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Cantador Ivan. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction (UMUAI)*, 26(Personality in Personalized Systems), 2016.

[119] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. Personality traits and music genres: What do people prefer to listen to? In *Proc. of the 25th Conf. on User Modeling, Adaptation and Personalization*, UMAP '17, pages 285–288, New York, NY, USA, 2017. ACM.

[120] Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, 55(5):697–712, 2009.

[121] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

[122] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[123] Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. Learning attribute-to-feature mappings for cold-start recommendations. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 176–185. IEEE, 2010.

[124] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41, 2016.

[125] Yue Gao, Meng Wang, Huanbo Luan, Jialie Shen, Shuicheng Yan, and Dacheng Tao. Tag-based social image search with visual-text joint hypergraph learning. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1517–1520. ACM, 2011.

[126] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260. ACM, 2010.

[127] Nuno Gil, Nuno Silva, Eduardo Duarte, Pedro Martins, Thibault Langlois, and Teresa Chambel. Going through the clouds: search overviews and browsing of movies. In *Proceeding of the 16th International Academic MindTrek Conference*, pages 158–165. ACM, 2012.

[128] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. *arXiv preprint arXiv:1704.02895*, 2017.

[129] Tobias Glasmachers. Limits of end-to-end learning. *arXiv preprint arXiv:1704.08305*, 2017.

[130] Rafael C Gonzalez and Richard E Woods. Digital image processing prentice hall. *Upper Saddle River, NJ*, 2002.

[131] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366, 2000.

[132] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(Dec):2935–2962, 2009.

[133] Asela Gunawardana and Guy Shani. Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer, 2015.

[134] Asela Gunawardana and Guy Shani. Evaluating Recommender Systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, chapter 8, pages 256–308. Springer, 2nd edition, 2015.

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

[136] Mohammad Haghighat, Mohamed Abdel-Mottaleb, and Wadee Alhalabi. Fully automatic face normalization and single sample face recognition in unconstrained environments. *Expert Systems with Applications*, 47:23–34, 2016.

[137] Michal Haindl. Texture synthesis. ERCIM, 1993.

[138] Donald E Hall. *Musical acoustics*. Brooks Cole, 2002.

[139] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.

[140] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[141] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[142] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

[143] David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalcic. The million musical tweets dataset: What can we learn from microblogs. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[144] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

Bibliography

[145] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: a visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 309–316. ACM, 2016.

[146] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 309–316, New York, NY, USA, 2016. ACM.

[147] Ruining He, Chunbin Lin, Jianguo Wang, and Julian McAuley. Sherlock: sparse hierarchical embeddings for visually-aware one-class collaborative filtering. *arXiv preprint arXiv:1604.05813*, 2016.

[148] Ruining He and Julian McAuley. 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*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

[149] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. 2016.

[150] Wolf-D Heine and Rainer Guski. Listening: the perception of auditory events? *Ecological Psychology*, 3(3):263–275, 1991.

[151] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[152] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[153] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical Symposium East*, pages 319–331. International Society for Optics and Photonics, 1981.

[154] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[155] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.

[156] Jing Huang, S Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997.

[157] Zan Huang, Hsinchun Chen, and Daniel Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):116–142, 2004.

[158] Ziwon Hyung, Kibeom Lee, and Kyogu Lee. Music recommendation using text analysis on song requests to radio stations. *Expert Systems with Applications*, 41(5):2608–2618, 2014.

[159] Bogdan Ionescu, Jan Schlüter, Ionut Mironica, and Markus Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrievalmm*, pages 215–222. ACM, 2013.

[160] Alice M Isen, Kimberly A Daubman, and Gary P Nowicki. Positive affect facilitates creative problem solving. *Journal of personality and social psychology*, 52(6):1122, 1987.

[161] Carroll E Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280, 2007.

[162] Dietmar Jannach, Lukas Lerche, and Iman Kamehkhosh. Beyond hitting the hits: Generating coherent music playlist continuations with the right tracks. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 187–194. ACM, 2015.

[163] Joris H Janssen, Egon L Van Den Broek, and Joyce HDM Westerink. Tune in to your emotions: a robust personalized affective music player. *User Modeling and User-Adapted Interaction*, 22(3):255–279, 2012.

[164] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[165] Rongrong Ji, Yue Gao, Richang Hong, Qiong Liu, Dacheng Tao, and Xuelong Li. Spectral-spatial constraint hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(3):1811–1824, 2014.

[166] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 27–34. ACM, 2015.

[167] Patrik N Juslin and John A Sloboda. *Music and emotion: Theory and research.* Oxford University Press, 2001.

[168] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112. ACM, 2013.

[169] Iman Kamehkhosh and Dietmar Jannach. User perception of next-track music recommendations. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 113–121. ACM, 2017.

[170] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.*, 7(1):2:1–2:42, December 2016.

[171] Marius Kaminskas and Francesco Ricci. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2):89–119, 2012.

[172] Marius Kaminskas, Francesco Ricci, and Markus Schedl. Location-aware music recommendation using auto-tagging and hybrid matching. In *Proceedings of the 7th ACM conf. on Recommender systems*, pages 17–24. ACM, 2013.

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

[174] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE, 2006.

[175] John Paul Kelly and Derek Bridge. Enhancing the diversity of conversational collaborative recommendations: a comparison. *Artificial Intelligence Review*, 25(1):79–95, Apr 2006.

[176] Patrick Kenny. A small footprint i-vector extractor. In *Odyssey*, volume 2012, pages 1–6, 2012.

[177] Hyon Hee Kim. A semantically enhanced tag-based music recommendation using emotion ontology. In *Asian Conference on Intelligent Information and Database Systems*, pages 119–128. Springer, 2013.

[178] Daniel Kluver and Joseph A Konstan. Evaluating recommender behavior for new users. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 121–128. ACM, 2014.

[179] Peter Knees and Markus Schedl. A Survey of Music Similarity and Recommendation from Music Context Data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 10(1), 2013.

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

[181] Peter Knees, Dominik Schnitzer, and Arthur Flexer. Improving neighborhood-based collaborative filtering by reducing hubness. In *Proceedings of International Conference on Multimedia Retrieval*, page 161. ACM, 2014.

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

[183] Silvia Knobloch. Mood adjustment via mass communication. *Journal of communication*, 53(2):233–250, 2003.

[184] Ron Kohavi. The power of decision tables. In *8th European Conference on Machine Learning*, pages 174–189. Springer, 1995.

[185] Ron Kohavi and Dan Sommerfield. Targeting business users with decision table classifiers. In *KDD*, pages 249–253, 1998.

[186] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016.

[187] Joseph A Konstan, Sean M McNee, Cai-Nicolas Ziegler, Roberto Torres, Nishikant Kapoor, and John Riedl. Lessons on applying automated recommender systems to information-seeking tasks. 2006.

[188] Joseph A Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.

[189] Irena Koprinska and Sergio Carrato. Temporal video segmentation: A survey. *Signal processing: Image communication*, 16(5):477–500, 2001.

[190] Yehuda Koren and Robert Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

[191] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42:30–37, August 2009.

[192] Andrej Košir, Ante Odic, Matevz Kunaver, Marko Tkalcic, and Jurij F Tasic. Database for contextual personalization. 2011.

[193] Bert Krages. *Photography: the art of composition*. Skyhorse Publishing, Inc., 2012.

[194] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[195] Anna M Kruspe. Improving singing language identification through i-vector extraction. In *DAFx*, pages 227–233, 2014.

[196] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 507–510. ACM, 2005.

[197] Paul Lamere. Social Tagging and Music Information Retrieval. *Journal of New Music Research: Special Issue: From Genres to Tags – Music Information Retrieval in the Age of Social Tagging*, 37(2):101–114, 2008.

[198] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1695–1699. IEEE, 2014.

[199] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.

[200] Michael S Lew. *Principles of visual information retrieval*. Springer Science & Business Media, 2013.

[201] Cheng-Te Li and Man-Kwan Shan. Emotion-based impressionism slideshow with automatic music accompaniment. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 839–842. ACM, 2007.

[202] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing*, 3(2):236–252, 2009.

[203] Tao Li, Mitsunori Ogihara, and Qi Li. A Comparative Study on Content-based Music Genre Classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR)*, Toronto, Canada, 2003.

[204] Wei Li, Farnaz Abtahi, and Zhigang Zhu. A deep feature based multi-kernel learning approach for video emotion recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 483–490. ACM, 2015.

[205] Yanen Li, Jia Hu, ChengXiang Zhai, and Ye Chen. Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 959–968. ACM, 2010.

[206] Bing Liu. *Web Data Mining – Exploring Hyperlinks, Contents and Usage Data*. Springer, Berlin, Heidelberg, Germany, 2007.

[207] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–360. ACM, 2009.

[208] Jing Liu, Zechao Li, Jinhui Tang, Yu Jiang, and Hanqing Lu. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Transactions on Multimedia*, 16(3):588–600, 2014.

[209] Ning-Han Liu. Comparison of content-based music recommendation using different distance estimation methods. *Applied intelligence*, 38(2):160–174, 2013.

[210] Wei Liu, Jun Wang, and Shih-Fu Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624–2638, 2012.

[211] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1):262–282, 2007.

[212] Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, MA, USA, 2000.

[213] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

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

[215] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1–49, 2012.

[216] Cui-Xia Ma, Yong-Jin Liu, Hong-An Wang, Dong-Xing Teng, and Guo-Zhong Dai. Sketch-based annotation and visualization in video authoring. *IEEE Transactions on Multimedia*, 14(4):1153–1165, 2012.

[217] Hao Ma, Jianke Zhu, Michael Rung-Tsong Lyu, and Irwin King. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12(5):462–473, 2010.

[218] Michael I. Mandel, Razvan Pascanu, Douglas Eck, Yoshua Bengio, Luca M. Aiello, Rossano Schifanella, and Filippo Menczer. Contextual Tag Inference. *ACM Transactions on Multimedia Computing, Communications and Applications*, 7S(1):32:1–32:18, 2011.

[219] Bangalore S Manjunath, Philippe Salembier, and Thomas Sikora. *Introduction to MPEG-7: multimedia content description interface*, volume 1. John Wiley & Sons, 2002.

[220] Gonçalo Marques, Thibault Langlois, Fabien Gouyon, Miguel Lopes, and Mohamed Sordo. Short-term feature space and music genre classification. *Journal of New Music Research*, 40(2):127–137, 2011.

[221] Oge Marques. *Practical image and video processing using MATLAB*. John Wiley & Sons, 2011.

[222] David Martınez, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka. Language recognition in ivectors space. *Proceedings of Interspeech, Firenze, Italy*, pages 861–864, 2011.

[223] Pascual Martínez-Gómez and Akiko Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 95–104. ACM, 2014.

[224] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-François Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[225] Y Matsuda. Color design. *Asakura Shoten*, 2(4):10, 1995.

[226] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[227] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*, pages 909–916. ACM, 2012.

[228] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101. ACM, 2006.

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

[230] Tao Mei, Bo Yang, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Shipeng Li. Videoreach: an online video recommendation system. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 767–768. ACM, 2007.

[231] Prem Melville, Raymod J. Mooney, and Ramadass Nagarajan. Content-boosted Collaborative Filtering for Improved Recommendations. In *Eighteenth National Conference on Artificial Intelligence*, pages 187–192, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[232] Robin L Nabi and Mary Beth Oliver. *The SAGE handbook of media processes and effects*. Sage, 2009.

[233] Frank Nack, Chitra Dorai, and S Venkatesh. Computational media aesthetics: Finding meaning beautiful. *IEEE multimedia*, 8(4):10–12, 2001.

[234] Alexandros Nanopoulos, Dimitrios Rafailidis, Panagiotis Symeonidis, and Yannis Manolopoulos. Musicbox: Personalized music recommendation based on cubic analysis of social tags. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):407–412, 2010.

[235] Xia Ning, Christian Desrosiers, and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook*, pages 37–76. Springer, 2015.

[236] Mark S Nixon and Alberto S Aguado. *Feature extraction & image processing for computer vision*. Academic Press, 2012.

[237] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver. The role of image composition in image aesthetics. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3185–3188. IEEE, 2010.

[238] Mary Beth Oliver. Mood management and selective exposure. 2003.

[239] Mary Beth Oliver. Tender affective states as predictors of entertainment preference. *Journal of Communication*, 58(1):40–61, 2008.

[240] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[241] Marco Paleari, Benoit Huet, and Ryad Chellali. Towards multimodal emotion recognition: a new approach. In *Proceedings of the ACM international conference on image and video retrieval*, pages 174–181. ACM, 2010.

[242] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based Organization and Visualization of Music Archives. In *Proceedings of the 10th ACM International Conference on Multimedia (MM 2002)*, pages 570–579, Juan les Pins, France, December 1–6 2002.

[243] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 570–579. ACM, 2002.

[244] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 653–658. ACM, 2004.

[245] Bo Peng, Lei Zhang, and David Zhang. A survey of graph theoretical approaches to image segmentation. *Pattern Recognition*, 46(3):1020–1038, 2013.

[246] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015.

[247] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[248] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[249] István Pilászy, Dávid Zibriczky, and Domonkos Tikk. Fast als-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 71–78. ACM, 2010.

[250] Konstantinos N Plataniotis and Anastasios N Venetsanopoulos. *Color image processing and applications*. Springer Science & Business Media, 2013.

[251] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[252] Martin F. Porter. *An Algorithm for Suffix Stripping*, pages 313–316. Morgan Kaufmann, San Francisco, CA, USA, 1997.

[253] Simon JD Prince and James H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[254] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164. ACM, 2011.

[255] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *arXiv preprint arXiv:1802.08452*, 2018.

[256] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010.

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

[258] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[259] J. Reed and C. Lee. Preference Music Ratings Prediction Using Tokenization and Minimum Classification Error Training. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2294–2303, 2011.

[260] W Scott Reilly. Believable social and emotional agents. Technical report, Carnegie-Mellon Univ Pittsburgh pa Dept of Computer Science, 1996.

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

[262] Steffen Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57, 2012.

[263] Steffen Rendle. Factorization machines with libFM. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57:1–57:22, May 2012.

[264] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.

[265] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 635–644. ACM, 2011.

[266] Seungmin Rho, Byeong-jun Han, and Eenjun Hwang. Svr-based music mood classification and context-based music recommendation. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 713–716. ACM, 2009.

[267] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 19–26, New York, NY, USA, 2012. ACM.

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

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

[270] Sujoy Roy and Sharath Chandra Guntuku. Latent factor representations for cold-start video recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 99–106, New York, NY, USA, 2016. ACM.

[271] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. Socialtransfer: cross-domain transfer learning from social streams for media applications. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 649–658. ACM, 2012.

[272] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[273] Thomas Schäfer and Claudia Mehlhorn. Can Personality Traits Predict Musical Style Preferences? A Meta-Analysis. *Personality and Individual Differences*, 116:265 – 273, 2017.

[274] János Schanda. *Colorimetry: understanding the CIE system*. John Wiley & Sons, 2007.

[275] Markus Schedl. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, New York, USA, June 2016.

[276] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41:523–539, December 2013.

[277] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Recsys challenge 2018: Automatic playlist continuation.

[278] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *IJMIR*, 7(2):95–116, 2018.

[279] Markus Schedl and Fang Zhou. Fusing Web and Audio Predictors to Localize the Origin of Music Pieces for Geospatial Retrieval. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*, Padua, Italy, March 2016.

[280] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, New York, NY, USA, 2002. ACM.

[281] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[282] Bill Schilit, Norman Adams, and Roy Want. Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE, 1994.

[283] Erik M. Schmidt and Youngmoo E. Kim. Projection of Acoustic Features to Continuous Valence-Arousal Mood Labels via Regression. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009.

[284] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, 13(Oct):2871–2902, 2012.

[285] Klaus Seyerlehner, Arthur Flexer, and Gerhard Widmer. On the limitations of browsing top-n recommender systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 321–324, New York, NY, USA, 2009. ACM.

[286] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner. A Refined Block-level Feature Set for Classification, Similarity and Tag Prediction. In *7th Annual Music Information Retrieval Evaluation eXchange (MIREX 2011)*, Miami, FL, USA, October 2011.

[287] Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees. Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation. In *Extended Abstract to the Music Information Retrieval Evaluation eXchange (MIREX 2010) / 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, the Netherlands, August 2010.

[288] Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter 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.

[289] Klaus Seyerlehner, Gerhard Widmer, and Tim Pohle. Fusing Block-Level Features for Music Similarity Estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 6-10 2010.

[290] Klaus Seyerlehner, Gerhard Widmer, and Tim Pohle. Fusing Block-Level Features for Music Similarity Estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.

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

[292] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. *Recommender systems handbook*, pages 257–297, 2011.

[293] Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara. Music recommendation based on acoustic features and user access patterns. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1602–1611, 2009.

[294] Yuesong Shen, Claire-Hélène Demarty, and Ngoc QK Duong. Technicolor @ mediaeval 2016 predicting media interestingness task. In *In Proc. of the MediaEval 2016 Workshop.*, 2016.

[295] Paul D Sherman. *Colour vision in the nineteenth century: the Young-Helmholtz-Maxwell theory*. Taylor & Francis, 1981.

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

[297] Yue Shi, Martha Larson, and Alan 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.

[298] Kijung Shin and U Kang. Distributed methods for high-dimensional and large-scale tensor factorization. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 989–994. IEEE, 2014.

[299] Gabriel S Simões, Jônatas Wehrmann, Rodrigo C Barros, and Duncan D Ruiz. Movie genre classification with convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint conf. on*, pages 259–266. IEEE, 2016.

[300] Barry Smyth and Paul McClave. Similarity vs. diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, ICCBR '01, pages 347–361, London, UK, 2001. SpringerVerlag.

[301] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM Multimedia*, 2005.

[302] Cees G. M. Snoek. *The Authoring Metaphor to Machine Understanding of Multimedia*. PhD thesis, University of Amsterdam, October 2005. You may obtain a bound booklet for free by sending me an e-mail.

[303] Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.

[304] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

[305] Yading Song, Simon Dixon, and Marcus Pearce. Evaluation of Musical Features for Emotion Classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012.

[306] Yading Song, Simon Dixon, and Marcus Pearce. A survey of music recommendation systems and future perspectives. In *9th International Symposium on Computer Music Modeling and Retrieval*, volume 4, 2012.

[307] Mohamed Sordo. *Semantic Annotation of Music Collections: A Computational Approach*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2012.

[308] Aleksandar Stupar and Sebastian Michel. Picasso-to sing, you must close your eyes and draw. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 715–724. ACM, 2011.

[309] Bob L. Sturm. A Survey of Evaluation in Music Genre Recognition. In Andreas Nürnberger, Sebastian Stober, Birger Larsen, and Marcin Detyniecki, editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, volume 8382 of *LNCS*. Springer, 2014.

[310] Jun-Won Suh, Seyed Omid Sadjadi, Gang Liu, Taufiq Hasan, Keith W Godin, and John HL Hansen. Exploring hilbert envelope based acoustic features in i-vector speaker verification using ht-plda. In *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, 2011.

[311] Lei Tang, Suju Rajan, and Vijay K Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of the 18th international conference on World wide web*, pages 211–220. ACM, 2009.

[312] Liang Tang, Yexi Jiang, Lei Li, and Tao Li. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 73–80. ACM, 2014.

[313] A Murat Tekalp. *Digital video processing*. Prentice Hall Press, 2015.

[314] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[315] Marco Tiemann and Steffen Pauws. Towards ensemble learning for hybrid music recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 177–178, New York, NY, USA, 2007. ACM.

[316] Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. Music playlist continuation by learning from hand-curated examples and song features: Alleviating the cold-start problem for rare and out-of-set songs. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, pages 46–54. ACM, 2017.

[317] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116. ACM, 2011.

[318] Saúl Vargas and Pablo Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys)*, Chicago, IL, USA, 2011.

[319] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[320] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, 2012.

[321] Fabio Vignoli and Steffen Pauws. A music retrieval system based on user driven similarity and its evaluation. In *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, pages 272–279, 2005.

[322] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[323] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. Computational analysis of sound scenes and events, 2018.

[324] Johann Wolfgang Von Goethe. *Theory of colours*, volume 3. Mit Press, 1840.

[325] Fanglin Wang, Daguang Li, and Mingliang Xu. A location-aware tv show recommendation with localized sementaic analysis. *Multimedia Systems*, 22(4):535–542, 2016.

[326] Hao Wang, Binyi Chen, and Wu-Jun Li. Collaborative topic regression with social regularization for tag recommendation.

[327] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, 2013.

[328] Diane Watson and Regan L Mandryk. An in-situ study of real-life listening context. *SMC 2012*, 2012.

[329] Felix Weninger, Florian Eyben, and Bjorn Schuller. On-line continuous-time music mood regression with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5412–5416. IEEE, 2014.

[330] Kris West and Paul Lamere. A model-based approach to constructing music similarity functions. *EURASIP Journal on Applied Signal Processing*, 2007(1):149–149, 2007.

[331] Rainer Westermann, GUNTER Stahl, and F Hesse. Relative effectiveness and validity of mood induction procedures: analysis. *European Journal of social psychology*, 26:557–580, 1996.

[332] Gin-Der Wu and Chin-Teng Lin. Word boundary detection with mel-scale frequency bank in noisy environment. *IEEE transactions on speech and audio processing*, 8(5):541–554, 2000.

[333] Rui Xia and Yang Liu. Using i-vector space model for emotion recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[334] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Video emotion recognition with transferred deep feature encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 15–22. ACM, 2016.

[335] Bin Xu, Jiajun Bu, Chun Chen, and Deng Cai. An exploration of improving collaborative recommender systems via user-item subgroups. In *Proceedings of the 21st international conference on World Wide Web*, pages 21–30. ACM, 2012.

[336] Songhua Xu, Hao Jiang, and Francis Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 83–90. ACM, 2008.

[337] Songhua Xu, Hao Jiang, and Francis Lau. User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 7–16. ACM, 2009.

[338] Songhua Xu, Yi Zhu, Hao Jiang, and Francis CM Lau. A user-oriented webpage ranking algorithm based on user attention time. 2008.

[339] Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing 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.

[340] Yi-Hsuan Yang and Homer H. Chen. Machine recognition of music emotion: A review. *Transactions on Intelligent Systems and Technology*, 3(3), May 2013.

[341] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *ISMIR*, volume 6, page 7th, 2006.

[342] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Improving efficiency and scalability of model-based music recommender system based on incremental training. In *ISMIR*, 2007.

[343] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):435–447, 2008.

[344] Herbert Zettl. *Sight, sound, motion: Applied media aesthetics*. Cengage Learning, 2013.

[345] Mi Zhang and Neil Hurly. Evaluating the diversity of top-n recommendations. In *Tools with Artificial Intelligence, 2009. ICTAI’09. 21st International Conference on*, pages 457–460. IEEE, 2009.

[346] Mi Zhang, Jie Tang, Xuchen Zhang, and Xiangyang Xue. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 73–82. ACM, 2014.

[347] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[348] Weinan Zhang, Li Tian, Xinruo Sun, Haofen Wang, and Yong Yu. A semantic approach to recommending text advertisements for images. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 179–186. ACM, 2012.

[349] Zhang, Yuan Cao and O Seaghdha, Diarmuid and Quercia, Daniele and Jambor, Tamas. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, Seattle, WA, USA, 2012.

Bibliography

[350] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1939–1946, 2014.

[351] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conf. on Multimedia*, pages 747–750. ACM, 2010.

[352] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 404–410. ACM, 2010.

[353] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[354] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on the World Wide Web*, pages 22–32. ACM, 2005.

[355] Dolf Zillmann. Mood management through communication choices. *American Behavioral Scientist*, 31(3):327–340, 1988.