Chapter 6
Enhancing Video Recommendation Using Multimedia Content

Yashar Deldjoo

Abstract Video recordings are complex media types. 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 modalities. To date, majority of movie recommender systems use collaborative filtering (CF) models or content-based filtering (CBF) relying on metadata (e.g., editorial such as genre or wisdom of the crowd such as user-generated tags) at their core since they are human-generated and are assumed to cover the 'content semantics' of movies by a great degree. The information obtained from multimedia content and learning from muli-modal sources (e.g., audio, visual and metadata) on the other hand, offers the possibility of uncovering relationships between modalities and obtaining an in-depth understanding of natural phenomena occurring in a video. These discerning characteristics of heterogeneous feature sets meet users' differing information needs. In the context of this Ph.D. thesis [9], which is briefly summarized in the current extended abstract, approaches to automated extraction of multimedia information from videos and their integration with video recommender systems have been elaborated, implemented, and analyzed. Variety of tasks related to movie recommendation using multimedia content have been studied. The results of this thesis can motivate the fact that recommender system research can benefit from knowledge in multimedia signal processing and machine learning established over the last decades for solving various recommendation tasks.

6.1 Introduction and Context

Users base their decision making about which movie to watch typically on its content, whether expressed in terms of metadata (e.g., genre, cast, or plot) or the feeling experienced after watching the corresponding movie trailer in which the visual content (e.g., color, lighting, motion) and the audio content (e.g., music or spoken dialogues)

Y. Deldjoo (✉)
SisInf Lab, Department of Electrical Engineering and Information Technology, Polytechnic University of Bari, Via Orabona, 4, 70125 Bari, Italy
e-mail: deldjooy@acm.org; yashar.deldjoo@poliba.it

B. Pernici (ed.), *Special Topics in Information Technology*, PoliMI SpringerBriefs,
https://doi.org/10.1007/978-3-030-32094-2_6

play a key role in users' perceived affinity to the movie. The above examples underline that human interpretation of media items is intrinsically content-oriented.

Recommender systems support users in their decision making by focusing them on a small selection of items out of a large catalogue. To date, most video recommendation models use collaborative filtering (CF), content-based filtering on metadata (CBF-metadata), or a combination thereof at their core [33, 35, 42]. While CF models exploit the correlations encoded in users' preference indicators—either implicit (clicks, purchases) or explicit (ratings, votes)—to predict the best-matching user-item pairs, CBF models use 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 compute recommendations. CBF approaches typically leverage *metadata* as a bridge between items and users, effectively disregarding a wealth of information encoded in the actual audio visual signals [13]. If we assume the primary role of a recommender system is to help people make choices that they will ultimately be satisfied with [7], such systems should thus take into account multiple source of information driving users' perception of media content and make a rational decision about their relative importance. This would in turn offer users the chance to learn more about their multimedia taste (e.g., their visual or musical taste) and their semantic interests [10, 13, 44].

The above are underlying ideas about why multimedia content can be useful for recommendation of *warm items* (videos with sufficient interactions). Nevertheless, in most video streaming services, new videos are continuously added. CF models are unable to make predictions in such scenario, since the newly added videos lack interactions, technically known as *cold-start* problem [5] and the associated items are referred by *cold items* (videos with few interactions) or *new items* (videos with no interactions). Furthermore, metadata can be rare/absent for cold/new videos, making it difficult to provide good quality recommendations [43]. Despite much research conducted in the field of RS for solving different tasks, the cold start (CS) problem is far from solved and most existing approaches suffer from it. Multimedia information *automatically extracted* from the audio-visual signals can serve as a proxy to solve the CS problem; in addition, it can act as a complementary information to identify videos that "look similar" or "sound similar" in warm start (WS) settings. These discerning characteristics of multimedia meet users' different information needs. As a branch of recommender systems, my Ph.D. thesis [9] 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.

6.2 Problem Formulation

In this section, we provide a formal definition of *content-based filtering video recommendation systems* (CBF-VRS) exploiting *multimedia content information*. In particular, we propose a general recommendation model of videos as composite

media objects, where the recommendation relies on the computation of distinct utilities, associated with image, audio, and textual modalities and a final utility, which is computed by aggregating individual utility values [22].[1]

A CBF-VRS based on multimedia content information is characterized by the following components:

1. Video Items: A video item s is represented by the triple: $s = (s_V, s_A, s_T)$ in which s_V, s_A, s_T refer to the *visual*, *aural*, and *textual* modalities, respectively. s_V encodes the *visual information* represented in the video frames; s_A encodes the *audio information* represented in sounds, music, spoken dialogues of a video; finally s_T is the metadata (e.g., genre labels, title) or natural language (e.g., sub-captions or speech spoken by humans and transcribed as text).

A video is a *multi-modal (composite)* media type (using s_A, s_V and s_T). This is while an audio item that represents performance of a classical music piece can be seen as an *uni-modal (atomic)* media type (using only s_A). A pop song with lyrics can be regarded as composite (using s_A and s_T), while an image of a scene or a silent movie atomic as well (using s_V). Multi-modal data supplies the system with rich and diverse information on the phenomenon relevant to the given task [27].

Definition 6.1 A CBF-VRS exploiting multimedia content is characterized as a system that is able to store and manage video items $s \in \mathcal{S}$, in which $\mathcal{S}$ is a repository of video items.

2. Multimedia Content-Based Representation: Developing a CBF-VRS based on multimedia content relies on content-based (CB) descriptions according to distinct modalities (s_V, s_A, s_T). From each modality, useful features can be extracted to describe the information of that modality. Different features can be classified based on several dimensions, e.g., the semantic expressiveness of features, level of granularity among others [4]. As for the former for instance, it is common to distinguish three levels of expressiveness, with increasing extent of semantic meaning: *low-level*, *mid-level*, and *high-level* features with respect to which features are categorized as shown in Table 6.1.

Over the last years, a large number of CB descriptors have been proposed to quantify various type of information in a video as summarized in Table 6.1. These descriptors are usually extracted by applying some form of signal processing or machine learning specific to a modality, and are described based on specific feature vectors. For example, in the visual domain, a rich suite of of low-level visual features are proposed by research in communities of multimedia, machine learning and computer vision for the purpose of image understanding, which we deem important for a CBF-VRS. The most basic and frequently used low-level features are *color*, *texture*, *edge* and *shape*, which are used to describe the "visual contents" of an image [26]. Besides, in the last two decades the need for devising descriptors that reduce or eliminate sensitivity to variations such as illumination, scale, rotation, and view point was

[1] Note that in this section, although definition of utilities are based on CBF model, in practice they can include a combination of CBF and CF models at their core.

Table 6.1 Categorization of different multimedia features based on their semantic expressiveness. Low-level features are close to the raw signal (e.g., energy of an audio signal, contrast in an image, motion in a video, or number of words in a text), while 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 between, mid-level features are more advanced than low-level ones, but farther away from being semantically meaningful as high-level ones. They are often expressed as a combination or transformations of low-level features, or they are inferred from low-level features via machine learning

Hierarchy/Modalities	Visual	Audio	Textual
High-level (semantic)	Events, story	Structure, mood, message	Story, writing style
Mid-level (syntactic)	Objects, people, their interaction	Note onsets, rhythm patterns	Sentence, term-frequency
Low-level (stylistic)	Motion, color, texture, shape	Pitch, timbre, loudness	Tokens, n-grams

recognized in the community of computer vision. This gave rise to the development of a number of popular computer vision algorithms for image understanding [46]. They include for instance scale invariant feature transform (*SIFT*) [36], speeded up robust features (*SURF*) [6], local binary patterns (*LBP*) [41], discrete Wavelet transform (*DWT*), such as Gabor filters [37], discrete Fourier transform (*DFT*), and histogram of oriented gradients (*HOG*) [8]. The peak of these developments was reached in the early 2010s, when deep convolutional neural networks (*CNNs*) achieved groundbreaking accuracy for image classification [34]. One of the most frequently stated advantages of the Deep Neural Networks (*DNNs*) is that, they leverage the representational power of high-level semantics encoded in DNNs to narrow the semantic gap between the visual contents of the image and high-level concepts in the user's mind when consuming media items.

Definition 6.2 A CBF-VRS exploiting multimedia content is a system that is able to process video items and represent each modality in terms of a feature vector $\mathbf{f_m} = [f_1,\ f_2,\ \ldots f_{|\mathbf{f_m}|}] \in \mathbb{R}^{|\mathbf{f_m}|}$ where $m \in \{V, A, T\}$ represents the visual, audio or textual modality.[2]

3. Recommendation Model: A recommendation model provides suggestions for items that are most likely of interest to a particular user [42].

Let $\mathcal{U}$ and $\mathcal{S}$ denote a set of users and items, respectively. Given a target user $u \in \mathcal{U}$, to whom the recommendation will be provided, and a repository of items $s \in \mathcal{S}$, the general task of a personalized recommendation model is to identify the video item s^* that satisfies

$$\forall u \in \mathcal{U}\ ,\ s_u^* = \arg\max_{s \in \mathcal{S}}\ R(u, s) \tag{6.1}$$

[2]Note that here we step our attention outside the end-to-end learning approaches often performed by deep neural networks where the intermediate step of feature extraction is not done explicitly, and instead feature extraction and the final machine learning task are jointly performed.

where $R(u, s)$ is the *estimated utility* of item s for the user u on the basis of which the items are ranked [3]. The utility is infact a measure of *usefulness* of an item to a user and is measured by the RS to judge how much an item is *worth* being recommended. For example, some examples of such a utility function include a utility represented by a *rating* or a *profit* function [1].

Definition 6.3 A **multi-modal CBF-VRS** is a system that aims to improve learning performance using the knowledge/information aquired from different data sources of different video modalities. The utility of recommendations in a multi-modal CBF-VRS can be specified with respect to several specific utilities computed across each modality, thus

$$\forall u \in \mathcal{U}\ ,\ s_u^* = \arg\max_{s \in \mathcal{S}}\ R(u, s) = F(R_m(u, s)) \tag{6.2}$$

where $R_m(u, s)$ denotes the utility of item s *for user* u with regards to modality $m \in \{V, A, T\}$, and F is an aggregation function of the estimated utilities for each modality.

Based on the semantics of the aggregation, different functions can be employed, each implying a particular interpretation of the affected process. A standard and simplest form of aggregation functions are *conjunctive* (such as min operator), *disjunctive* (such as max operator), and *averaging* [39, 45]. As an example of the latter, and the one used in the field of multimedia information retrieval (MMIR), the weighted average linear combination is commonly used thus

$$R(u, s) = \sum_m w_m R_m(u, s) \tag{6.3}$$

where w_m is a weight factor indicating the importance of modality m, known as modality weights. The weights can be chosen as fixed weights or learned via machine learning. For example, recently studies based on *dictionary learning*, *co-clustering* and *multi-modal topic modeling* have become increasingly popular paradigms for the task of multi-modal inference [30]. For instance, multi-modal topic modelling (commonly methods based on latent semantic analysis or latent Dirichlet allocation) [30] models visual, audio and textual words with an underlying latent topic space.

6.3 Brief Overview of Ph.D. Research

The main processing stages involved in a CBF-VRS exploiting multimedia content are shown in Fig. 6.1. The input information are videos (movies) and the preference indications of a single user on them, and the output is a rank list of recommended videos (movies) tailored to target user's preference on the content

- **Temporal segmentation**: The goal of temporal segmentation is to partition the video item—infact the audio and image signals—into smaller structural units that

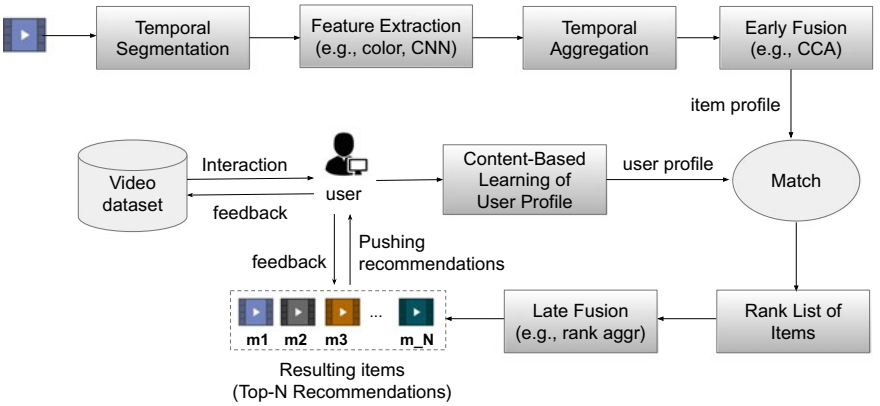

Fig. 6.1 The general framework illustrating the main processing steps involved in a CBF-VRS exploiting multimedia content. The framework focuses on the multimedia processing stages required to build a CBF system, however it can be extended to incorporate CF knowledge (e.g., in a hybrid CBF+CF system) and contextual factor (e.g., in context-aware CBF system)

have similar semantic content [29]. For the audio domain, segmentation approaches commonly operate at the *frame-level* or the *block-level*, the latter sometimes referred to as *segment-level* [25, 31]. For the visual domain, temporal segmentation segments the video into shots based on visual similarity between consecutive frames [15]. Some works consider scene-based segmentation, where a scene is semantically/hierarchically a higher-level video unit compared to a shot [15, 32]. Yet a simpler approach relies on capturing video at a fixed frame rate e.g., 1 fps and use all the resulting frames for processing [40].

- **Feature Extraction**: Feature extraction algorithms aim at encoding the content of the multimedia items in a concise and descriptive way, so to represent them for further use in retrieval, recommendation or similar systems [38]. An accurate feature representation can reflect item characteristics from various perspectives and can be highly indicative of user preferences.
  In the context of this Ph.D., a wide set of audio and visual features has been used to solve different movie recommendation tasks. As for the visual domain, they include *mise-en-scène visual features* (average short length, color variation, motion and lighting key) [12, 15], visual features based on the *MPEG-7 standard* and *pre-trained CNNs* [17] and the most stable and recent datasets, named Multifaceted Movie Trailer Feature dataset (MMTF-14K) and Multifaceted Video Clip Dataset (MVCD-7K) [11, 21], which we made publicly available online. In particular, MMTF-14K[3] provides state-of-the-art audio and visual descriptors for approximately 14K Hollywood-type *movie trailers* accompanied with metadata and user preference indicators on movies that are linked to the ML-20M dataset. The visual descriptors consist of two categories of descriptors: *aesthetic features*

[3] https://mmprj.github.io/mtrm_dataset/index.

and *pre-trained CNN* (AlexNet) features, each of them including different aggregation schemes for the two types of visual features. The audio descriptors consist of two classes of descriptors: *block-level features* (BLF) and *i-vector features* capturing spectral and timbral features of audio signal. To the best of our knowledge, MMTF-14K is the only large scale **multi-modal** dataset to date providing a rich source for devising and evaluating movie recommender systems.

A criticism of the MMTF-14K dataset however is that its underlying assumption relies on the fact that *movie trailers are representative of full movies*. Movie trailers are human-edited and artificially produced with lots of thrills and chills as their main goal is to motivate users to come back (to the cinema) and watch the movie. For this reason, the scenes in trailers are usually taken from the most exciting, funny, or otherwise noteworthy parts of the film,[4] which is a strong argument against the representativeness of trailers for the full movie. To address these shortcomings, in 2019 we introduced a new dataset of video clips, named Multi-faceted Video Clip Dataset (MFVCD-7K).[5] Each movie in MFVCD-7K can have several associated video clips, each focused on a particular scene, displaying it at its natural pace. Thus, video clips in MFVCD-7K can serve as a more *realistic* summary of the movie story than trailers.

- **Temporal aggregation**: This step involves creating a video-level descriptor by aggregating the features temporally. The following approaches are widely used in the field of multimedia processing: (i) *statistical summarization*: it is the simplest approach using the operators mean, standard deviation, median, maximum, or combinations thereof, e.g., means plus covariance matrix to build an item-level descriptor; (ii) *probabilistic modeling*: it is an alternative approach for temporal aggregation, which summarizes the local features of the item under consideration by a probabilistic model. Gaussian mixture models (GMMs) are often used for this purpose; (iii) *other approaches*: other feature aggregation techniques include vector quantization (VQ), vectors of locally aggregated descriptors (VLAD) and Fisher vectors (FV), where the last two were originally used for aggregating image key-point descriptors. They are used as a post-processing step for video representation, for example within a convolutional neural network (CNN) [28]. For different movie recommendation tasks, we used different temporal aggregation functions [12, 13, 15].
- **Fusion**: This step is the main step toward building a multi-modal VRS. Early fusion attempts to combine feature extracted from various unimodal streams into a single representation. For instance, in [13, 16] we studied adoption of an effective early fusion technique named *canonical correlation analysis* (CCA) to combine visual, textual and/or audio descriptors extracted from movie trailers and better exploit complementary information between different modalities. Late fusion approaches combine outputs of several system run on different descriptors. As an example of this approach, in [11, 13], we used a novel late fusion strategy based on a weighted variant of the Borda rank aggregation strategy to combine heterogeneous feature

[4] https://filmshortage.com/the-art-of-the-trailer/.

[5] https://mmprj.github.io/MFVCD-7K.

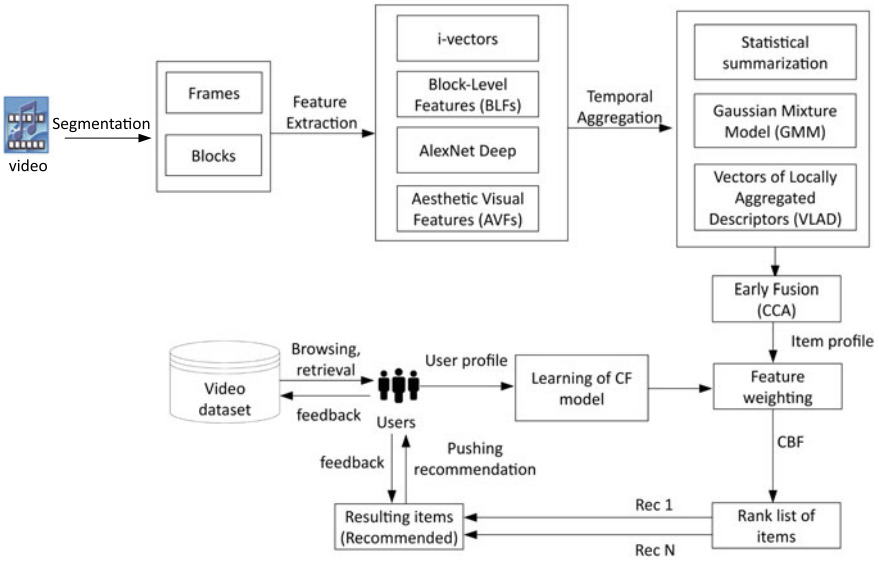

Fig. 6.2 The proposed collaborative-filtering-enriched content-based filtering (CFeCBF) movie recommender system framework proposed in [13] to solve the new/cold-item problem

sets into a unified ranking of videos and showed promising improvements of the final ranking. Note that a different level of hybridization—from a recommendation point of view—can involve fusing CBF system with a CF model, which we considered e.g., in [14, 17].

- **Content-based learning of user profile and Recommendation**: The goal of this step is to learn a user-specific model which is used to predict the target user's interest in (multimedia) items based on her past history of interaction with the items [2]. The learned user profile model is compared to representative item features (or item profiles) in order to make recommendations tailored to target user's preference on the content.

For instance, in [10] a *multi-modal content-based movie recommender system* is presented that exploits rich content descriptors based on state-of-the-art multimedia descriptors: *block-level and i-vector features for audio* and *aesthetic and deep visual features*. For multi-modal learning, a novel late fusion strategy based on an extended version of the Borda rank aggregation strategy was proposed which resulted in an improved ranking of videos. Evaluation was carried out on a subset of MovieLens-20M and multimedia features extracted from 4,000 movie trailers, by (i) a *system-centric study* to measure the offline quality of recommendations in terms of accuracy-related (MRR, MAP, recall) and beyond-accuracy (novelty, diversity, coverage) performance, and (ii) a *user-centric online experiment*, measuring different subjective metrics (relevance, satisfaction, diversity). Results of empirical evalua-

Table 6.2 Comparison of different research works carried out in the context of this Ph.D. thesis. Abbreviations: Focus: Focus of Study, WS: Warm Start, CS: Cold Start, Meta Type: Metadata Type, Ed: Editorial Metadata (e.g., genre labels), UG: User-generated metadata (e.g., tags), M-mod Fusion: Multi-modality Fusion Type, Rec Model: Recommendation Model, CF: Collaborative Filtering, CBF: Content-based Filtering, CA: Context-Aware, Acc: Accuracy Metric (e.g., MAP, NDCG), Beyond: Beyond Accuracy Metric (novelty, diversity, coverage)

Res	Year	Focus		Content modality			Meta type		M-mod fusion		Rec model			Eval type		Eval metric	
		WS	CS	Audio	Visual	Textual	Ed	UG	Early	Mid/Late	CBF	CF	CA	Offline	User-study	Acc	Beyond
[18]	2015	✓			✓						✓			✓		✓	
[15]	2016	✓			✓						✓			✓		✓	
[14]	2016	✓			✓						✓	✓		✓		✓	
[12]	2017	✓			✓						✓			✓		✓	
[24]	2017	✓			✓	✓	✓			✓	✓			✓	✓	✓	✓
[19]	2017	✓			✓								✓				
[20]	2017	✓			✓								✓				
[10]	2018	✓		✓	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓
[11]	2018		✓	✓	✓	✓	✓	✓		✓	✓			✓		✓	
[17]	2018	✓			✓		✓		✓		✓	✓		✓		✓	
[13]	2019	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓

tion indicates that multimedia features can provide a good alternative to metadata (as baseline), with regards to both accuracy measures and beyond accuracy measures.

In [13] a novel *movie multi-modal recommender system* is proposed that specifically addresses the *new item cold-start problem* by: (i) integrating state-of-the-art audio and visual descriptors, which can be automatically extracted from video content and constitute what we call the *movie genome*; (ii) exploiting an effective data fusion method named *canonical correlation analysis* (CCA) to better exploit complementary information between different modalities; (iii) proposing a two-step hybrid approach which trains a CF model on warm items (items with interactions) and leverages the learned model on the movie genome to recommend cold items (items without interactions). The recommendation method is thus named collaborative-filtering enriched CBF (CFeCBF), which has a different functioning concept compared with a standard CBF system (compare Figs. 6.1 and 6.2). Experimental validation is carried out using a system-centric study on a large-scale, real-world movie recommendation dataset both in an absolute cold start and in a cold to warm transition; and a user-centric online experiment measuring different subjective aspects, such as satisfaction and diversity. Results from both the offline study as well as a preliminary user-study confirm the usefulness of their model for new item cold start situations over current editorial metadata (e.g., genre and cast).

Finally, in Table 6.2, we provide a brief comparison of a selected number of research works completed in the course of this Ph.D. thesis by highlighting their main aspects. Readers are referred to the comprehensive literature review on recommender system leveraging multimedia content in which I describe many domains where multimedia content plays a key role in human decision making and are considered in the recommendation process [23].

6.4 Conclusion

This extended abstract briefly discusses the main outcomes of my Ph.D. thesis [9]. This Ph.D. thesis studies video recommender systems using multimedia content in detail—a particular area in the design space of recommender system algorithms where the generic recommender algorithm can be configured in order to integrate a rich source of information extracted from the actual audio-visual signals of video. I believe different systems, techniques and tasks for movie recommendation, which were studied in this Ph.D. thesis can pave the path for a new paradigm of video (and in general multimedia) recommender system by designing recommendation models built on top of rich item descriptors extracted from content.

References

1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749. https://doi.org/10.1109/TKDE.2005.99
2. Aggarwal CC (2016) Content-based recommender systems. Recommender systems. Springer, Berlin, pp 139–166
3. Aggarwal CC (2016) An introduction to recommender systems. Recommender systems. Springer, Berlin, pp 1–28
4. Al-Halah Z, Stiefelhagen R, Grauman K (2017) Fashion forward: forecasting visual style in fashion. In: IEEE international conference on computer vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp 388–397. https://doi.org/10.1109/ICCV.2017.50.
5. Asmaa Elbadrawy GK (2015) User-specific feature-based similarity models for top-n recommendation of new items. ACM Trans Intell Syst, 6. https://doi.org/10.1145/2700495
6. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision, pp 404–417. Springer
7. Chen L, De Gemmis M, Felfernig A, Lops P, Ricci F, Semeraro G (2013) Human decision making and recommender systems. ACM Trans Interact Intell Syst (TiiS) 3(3):17
8. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, CVPR 2005, vol 1, pp 886–893. IEEE
9. Deldjoo Y (2018) Video recommendation by exploiting the multimedia content. PhD thesis, Italy
10. Deldjoo Y, Constantin MG, Eghbal-Zadeh H, Ionescu B, Schedl M, Cremonesi P (2018) Audio-visual encoding of multimedia content for enhancing movie recommendations. In: Proceedings of the 12th ACM conference on recommender systems, RecSys 2018, Vancouver, BC, Canada, October 2–7, 2018, pp 455–459. https://doi.org/10.1145/3240323.3240407
11. Deldjoo Y, Constantin MG, Ionescu B, Schedl M, Cremonesi P (2018) MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In: Proceedings of the 9th ACM multimedia systems conference, MMSys 2018, Amsterdam, The Netherlands, June 12–15, 2018, pp 450–455. https://doi.org/10.1145/3204949.3208141
12. Deldjoo Y, Cremonesi P, Schedl M, Quadrana M (2017) The effect of different video summarization models on the quality of video recommendation based on low-level visual features. In: Proceedings of the 15th international workshop on content-based multimedia indexing, CBMI 2017, Florence, Italy, June 19–21, 2017, pp 20:1–20:6. https://doi.org/10.1145/3095713.3095734
13. Deldjoo Y, Dacrema MF, Constantin MG, Eghbal-zadeh H, Cereda S, Schedl M, Ionescu B, Cremonesi P (2019) Movie genome: alleviating new item cold start in movie recommendation. User Model User-Adapt Interact 29(2):291–343. https://doi.org/10.1007/s11257-019-09221-y
14. Deldjoo Y, Elahi M, Cremonesi P (2016) Using visual features and latent factors for movie recommendation. In: Proceedings of the 3rd workshop on new trends in content-based recommender systems co-located with ACM conference on recommender systems (RecSys 2016), Boston, MA, USA, September 16, 2016, pp 15–18. http://ceur-ws.org/Vol-1673/paper3.pdf
15. Deldjoo Y, Elahi M, Cremonesi P, Garzotto F, Piazzolla P, Quadrana M (2016) Content-based video recommendation system based on stylistic visual features. J Data Semant 5(2):99–113. https://doi.org/10.1007/s13740-016-0060-9
16. Deldjoo Y, Elahi M, Cremonesi P, Moghaddam FB, Caielli ALE (2016) How to combine visual features with tags to improve movie recommendation accuracy? In: International conference on electronic commerce and web technologies, pp 34–45. Springer
17. Deldjoo Y, Elahi M, Quadrana M, Cremonesi P (2018) Using visual features based on MPEG-7 and deep learning for movie recommendation. IJMIR 7(4):207–219. https://doi.org/10.1007/s13735-018-0155-1

18. Deldjoo Y, Elahi M, Quadrana M, Cremonesi P, Garzotto F (2015) Toward effective movie recommendations based on mise-en-scène film styles. In: Proceedings of the 11th biannual conference on Italian SIGCHI chapter, CHItaly 2015, Rome, Italy, September 28–30, 2015, pp 162–165. https://doi.org/10.1145/2808435.2808460
19. Deldjoo Y, Frà C, Valla M, Cremonesi P (2017) Letting users assist what to watch: an interactive query-by-example movie recommendation system. In: Proceedings of the 8th Italian information retrieval workshop, Lugano, Switzerland, June 05–07, 2017, pp 63–66. http://ceur-ws.org/Vol-1911/10.pdf
20. Deldjoo Y, Frà C, Valla M, Paladini A, Anghileri D, Tuncil MA, Garzotta F, Cremonesi P et al (2017) Enhancing children's experience with recommendation systems. In: Workshop on children and recommender systems (KidRec'17)-11th ACM conference of recommender systems, pp N–A
21. Deldjoo Y, Schedl M (2019) Retrieving relevant and diverse movie clips using the mfvcd-7k multifaceted video clip dataset. In: Proceedings of the 17th international workshop on content-based multimedia indexing
22. Deldjoo Y, Schedl M, Cremonesi P, Pasi G (2018) Content-based multimedia recommendation systems: definition and application domains. In: Proceedings of the 9th Italian information retrieval workshop, Rome, Italy, May, 28–30, 2018. http://ceur-ws.org/Vol-2140/paper15.pdf
23. Deldjoo Y, Schedl M, Cremonesi P, Pasi G (2020) Recommender systems leveraging multimedia content. ACM Comput Surv (CSUR)
24. Elahi M, Deldjoo Y, Moghaddam FB, Cella L, Cereda S, Cremonesi P (2017) Exploring the semantic gap for movie recommendations. In: Proceedings of the Eleventh ACM conference on recommender systems, RecSys 2017, Como, Italy, August 27–31, 2017, pp 326–330. https://doi.org/10.1145/3109859.3109908
25. Ellis DP (2007) Classifying music audio with timbral and chroma features. ISMIR 7:339–340
26. Flickner M, Sawhney HS, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P (1995) Query by image and video content: the QBIC system. IEEE Comput 28(9):23–32. https://doi.org/10.1109/2.410146
27. Geng X, Wu X, Zhang L, Yang Q, Liu Y, Ye J (2019) Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting. arXiv:1905.11395
28. Girdhar R, Ramanan D, Gupta A, Sivic J, Russell B (2017) Actionvlad: Learning spatio-temporal aggregation for action classification. arXiv:1704.02895
29. Hu W, Xie N, Li L, Zeng X (2011) Maybank S (2011) A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern Part C (Applications and Reviews) 41(6):797–819
30. Irie G, Liu D, Li Z, Chang S (2013) A bayesian approach to multimodal visual dictionary learning. In: 2013 IEEE conference on computer vision and pattern recognition, Portland, OR, USA, June 23–28, 2013, pp 329–336. https://doi.org/10.1109/CVPR.2013.49
31. Knees P, Schedl M (2013) A survey of music similarity and recommendation from music context data. ACM Trans Multimed Comput Commun Appl (TOMCCAP) 10(1)
32. Koprinska I, Carrato S (2001) Temporal video segmentation: a survey. Signal Process Image Commun 16(5):477–500
33. Koren Y, Bell R (2015) Advances in collaborative filtering. In: Recommender systems handbook, pp 77–118. Springer
34. Liu L, Chen J, Fieguth P, Zhao G, Chellappa R, Pietikainen M (2018) A survey of recent advances in texture representation. arXiv:1801.10324
35. Lops P, De Gemmis M, Semeraro G (2011) Content-based recommender systems: state of the art and trends. In: Recommender systems handbook, pp 73–105. Springer, Berlin
36. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
37. Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 18(8):837–842
38. Marques O (2011) Practical image and video processing using MATLAB. Wiley, New York

39. Marrara S, Pasi G, Viviani M (2017) Aggregation operators in information retrieval. Fuzzy Sets Syst 324:3–19
40. Ng JY, Hausknecht MJ, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101
41. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
42. Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Recommender systems handbook, pp 1–34. Springer, Berlin
43. Roy S, Guntuku SC (2016) Latent factor representations for cold-start video recommendation. In: Proceedings of the 10th ACM conference on recommender systems, pp 99–106. ACM
44. Swearingen K, Sinha R (2002) Interaction design for recommender systems. Des Interact Syst 6:312–334
45. Tzeng GH, Huang JJ (2011) Multiple attribute decision making: methods and applications. CRC Press, Boca Raton
46. Vedaldi A, Fulkerson B (2008) VLFeat: an open and portable library of computer vision algorithms. http://www.vlfeat.org/