Noname manuscript No.
(will be inserted by the editor)

Content-based Video Recommendation System based on Stylistic Visual Features

Yashar Deldjoo · Mehdi Elahi · Paolo Cremonesi · Franca Garzotto · Pietro Piazzolla · Massimo Quadrana

Received: date / Accepted: date

Abstract This paper investigates the use of automatically extracted visual features of videos in the context of recommender systems and brings some novel contributions in the domain of video recommendations. We propose a new content-based recommender system that encompasses a technique to automatically analyze video contents and to extract a set of representative stylistic features (lighting, color, and motion) grounded on existing approaches of Applied Media Theory.

The evaluation of the proposed recommendations, assessed w.r.t. relevance metrics (e.g., recall) and compared with existing content-based recommender systems that exploit explicit features such as movie genre, shows that our technique leads to more accurate recommendations. Our proposed technique achieves better results not only when visual features are extracted from full-length videos, but also when the feature extraction technique operates on movie trailers, pinpointing that our approach is effective also when full-length videos are not available or when there are performance requirements.

Our recommender can be used in combination with more traditional content-based recommendation techniques that exploit explicit content features associated to video files, in order to improve the accuracy of recommendations. Our recommender can also be used alone, to address the problem originated from video files that have no meta-data, a typical situation of popular movie-sharing websites (e.g., YouTube) where every day hundred millions of hours of videos are uploaded by users and may contain no associated information. As they lack explicit content, these items cannot be considered for recommendation purposes by conventional content-based techniques even when they could be relevant for the user.

Yashar Deldjoo
Politecnico di Milano
Milan, Italy
Italia
E-mail: yashar.deldjoo@polimi.it

Mehdi Elahi
Politecnico di Milano
Milan, Italy
E-mail: mehdi.elahi@polimi.it

Paolo Cremonesi
Politecnico di Milano
Milan, Italy
E-mail: paolo.cremonesi@polimi.it

Franca Garzotto
Politecnico di Milano
Milan, Italy
E-mail: franca.garzotto@polimi.it

Pietro Piazzolla
Politecnico di Milano
Milan, Italy
E-mail: pietro.piazzolla@polimi.it

Massimo Quadrana
Politecnico di Milano
Milan, Italy
E-mail: massimo.quadrana@polimi.it

1 Introduction

Recommender Systems (RSs) are characterized by the capability of filtering large information spaces and selecting the items that are likely to be more interesting and attractive to a user [44]. Recommendation methods are usually classified into collaborative filtering methods, content-based methods and hybrid methods [3, 13, 44, 47]. Content-based methods, that are among popular ones [36, 5, 41], suggest items which have content

characteristics similar to the ones of items a user liked in the past. For example, news recommendations consider words or terms in articles to find similarities.

A prerequisite for content-based filtering is the availability of information about relevant content features of the items. In most existing systems, such features are associated to the items as structured or un-structured meta-information. Many RSs in the movie domain, for instance, consider movie genre, director, cast, (structured information), or plot, tags and textual reviews (un-structured 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 content-based technique that filters items according to visual features extracted automatically from video files, either full-length videos or trailers. Such features include lighting, color, and motion; they have a "stylistic" nature and, according to Applied Media Aesthetics [53], can be used to convey communication effects and to stimulate different feelings in the viewers.

The proposed recommendation technique has been evaluated w.r.t. relevance metrics (e.g., recall), using conventional techniques for the off-line evaluation of recommender systems that exploit machine learning methods [22,17]. The results have been then compared with existing content-based 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 videos; and (c) traditional explicit features based on genre – in order to test two hypotheses:

1. Our recommendation algorithm based on visual features leads to a higher recommendation accuracy in comparison with conventional genre-based recommender systems.
2. Accuracy is higher when stylistic features are extracted from either full-length movies or when they originate from movie trailers only. In other words, for our recommender movie trailers are good representatives of their corresponding full-length movies.

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.

Our work provides a number of contributions to the RS field in the video domain. It improves our understanding on the role of implicit visual features in the recommendation process, a subject which has been addressed by a limited number of researches. The proposed technique can be used in two ways:

- "In combination with" other content-based techniques that exploit explicit content, in order to improve their accuracy. This mixed approach has been investigated and evaluated by other works [52,54]. Still, prior off-line evaluations have involved a limited number of users (few dozens) against the thousands employed in our study.
- "Autonomously", to replace traditional content-based approaches when (some) video items (typically the new ones) are not equipped with the explicit content features that a conventional recommender would employ to generate relevant recommendations. This situation, which hereinafter we refer to as "*extreme new item problem*" [25] typically occurs for example in popular movie-sharing websites (e.g., YouTube) where every day hundred millions of hours of videos are uploaded by users and may contain no metadata. Conventional content-based techniques would neglect to consider these new items even if they may be relevant for recommendation purposes, as the recommender has no content to analyze but video files. To our knowledge, the generation of recommendations that exploit automatically extracted visual features "only" has not been explored nor evaluated in prior works.

As an additional contribution, our study pinpoints that our technique is accurate when visual feature extraction operates both on full-length movies (which is a computationally demanding process) and on movie trailers. Hence our method can be used effectively also when full-length videos are not available or when it is important to improve performance.

The rest of the paper is organized as follows. Section 2 reviews the relevant state of the art, related to content-based recommender systems and video recommender systems. This Section also introduces some theoretical background on Media Aesthetics that helps us to motivate our approach and interpret the results of our study. Section 3 describes the possible relation between the visual features adopted in our work and the aesthetic variables that are well known for artists in the domain of movie making. In Section 4 we describe our method for extracting and representing implicit visual features of the video and provide the details of our recommendation algorithm. Section 5 introduces the evaluation method. Section 6 presents the results of the study and Section 7 discusses them. Section 8 draws the conclusions and identifies open issues and directions for future work.

2 Related work

2.1 Content-Based Recommender Systems

Content-based RSs create a profile of a user's preferences, interests and tastes by considering the feedback provided by the user to some items together with the content associated to them. Feedback can be gathered either explicitly from users, by explicitly asking them to rate an item [7], or implicitly by analyzing her activity [30]. Recommendations are then generated by matching the user profile against the features of all items. Content can be represented using keyword-based models, in which the recommender creates a Vector Space Model (VSM) representation of item features, where an item is represented by a vector in a multi-dimensional space. These dimensions represent the features used to describe the items. By means of this representation, the system measures a relevance score that represents the user's degree of interest toward any of these items [36]. For instance, in the movie domain, the features that describe an item can be genre, actors, or director [40]. This model may allow content-based recommender systems to naturally tackle the new item problem [25]. Other families of content-based RSs use semantic analysis (lexicons and ontologies) to create more accurate item representations [19, 24, 37].

In the literature, a variety of content-based recommendation algorithms have been proposed. A traditional example is the "k-nearest neighbor" approach (KNN) that computes the preference of a user for an unknown item by comparing it against all the items known by the user in the catalogue. Every known item contributes to predict the preference score according to its similarity with the unknown item. The similarity can be measured by typically using Cosine similarity [36, 41]. There are also works that model the probability for the user to be interested to an item using a Bayesian approach [39], or use other techniques adopted from IR (Information Retrieval) such as the Relevance Feedback method [4].

2.2 Recommender Systems in the multimedia domain

In the multimedia domain, recommender systems typically exploit two types of item features, hereinafter referred to as High-Level features(HL) or Low-Level features(LL). High-Level features express properties of media content that are obtained from structured sources of meta-information such as databases, lexicons and ontologies, or from less structured data such as reviews, news articles, item descriptions and social tags [4, 15, 8, 38–40]. Low-Level features are extracted directly from media files themselves [20, 21]. In the music recommendation domain, for example, Low-Level features are acoustic properties such as rhythm or timbre, which are exploited to find music tracks that are similar to those liked by a user [9, 10, 31, 46].

In the domain of video recommendation, a limited number of works have investigated the use of Low-Level features, extracted from pure visual contents, which typically represent stylistic aspect of the videos [34, 52, 54, 14]. Still, existing approaches consider only scenarios where Low-Level features are exploited in addition to another type of information with the purpose of improving the quality of recommendations. [52] proposes a video recommender system, called VideoReach, which use a combination of High-Level and Low-Level video features of different nature - textual, visual and aural - to improve the click-through-rate. [54] proposes a multi-task learning algorithm to integrate multiple ranking lists, generated by using different sources of data, including visual content. As none of these works use Low-Level visual features only, they cannot be applied when the extreme new item problem [45] occurs, i.e., when only video files are available and high-level information is missing.

2.3 Video Retrieval

A Content-Based Recommender System (CBRS) for videos is similar to a Content-Based Video Retrieval system (CBVR) in the sense that both systems analyze video content in order to search for digital videos in large video databases. However, there are major differences between these systems. For example, people use popular video sharing websites such as YouTube for three main purposes [18]: (i) Direct navigation: to watch videos that they found at specific websites, (ii) Search: to watch videos around a specific topic expressed by a set of keywords, (iii) Personal entertainment: to be entertained by the content that matches their taste. A CBVR system is composed of a set of techniques that typically address the first and the second goal, while a CBRS focuses on the third goal. Accordingly, the main differences between CBRS and CBVR can be listed as follows [52]:

1. Different objectives: The goal of a CBVR system is to search for videos that match "a given query" provided directly by a user as a textual or video query *etc.* The goal of a video CBRS is however to search for videos that are matched with "user taste" (also known as user profile) and can be obtained by analyzing his past behavior and opinions on different videos.

2. Different inputs: The input to a CBVR system typically consists of a set of keywords or a video query where the inputs could be entirely unstructured and do not have any property per se. The input to a video CBRS on top of the video content includes some or many features obtained from user modeling (user profile, tasks, activities), the context (location, time, group) and other sources of information.
3. Different features: In general, video content features can be classified into 3 rough hierarchical levels [51]:
   - Level 1: Stylistic low-level that deals with modeling the visual styles in a video.
   - Level 2: Syntactic level that deals with finding objects and their interaction in a video.
   - Level 3: Semantic level that deals with conceptual modeling of a video.

   People most often rely on content features derived from level 2 and 3 in order to search for videos as they reside closer to human understanding. Even for recommender systems, most CBRSs use video metadata (genre, actor etc.) that reside in higher syntactic and semantical levels in order to provide recommendations. One of the novelty of this work is to explore the importance of stylistic low-level features in human's perception of movies. Movie directors drastically use the human's perception in stages of movie creation, in order to convey emotions and feeling to the users. We thus conclude that CBVR system and CBRS deal with video content modeling at different levels depending on suitability of the feature for a particular application.

While low-level features have been marginally explored in the community of recommender systems, they have been extensively studied in other fields such as Computer Vision and Content-Based Video Retrieval [43, 35, 29]. Although for different objectives, these communities share with the community of recommender systems, the research problems of defining the "best" representation of video content and of classifying videos according to features of different nature. Hence they offer results and insights that are of interest also in the movie recommender systems context. [29, 11] provide comprehensive surveys on the relevant state of the art related to video content analysis and classification, and discuss a large body of low-level features (visual, auditory or textual) that can be considered for these purposes. In [43] Rasheed et al. proposes a practical movie genre classification scheme based on computable visual cues. [42] discusses a similar approach by considering also the audio features. Finally, in [55] Zhou et al. proposes a framework for automatic classification, using a temporally-structured features, based on the intermediate level of scene representation.

2.4 Video features from a semiotic perspective

The stylistic visual features of videos that we exploit in our recommendation algorithm have been studied not only in Computer Science but also from a semiotic and expressive point of view, in the theory and practice of movie making (see Section 3). Lighting, color, and camera motion are important elements that movie makers consider in their work in order to convey meanings, or achieve intended emotional, aesthetic, or informative effects. Applied Media Aesthetic [53] is explicitly concerned with the relation of media features (e.g., lights, shadows, colors, space representation, camera motion, or sound) with perceptual, cognitive, and emotional reactions they are able to evoke in media consumers, and tries to identify patterns in how such features operate to produce the desired effect [23]. Some aspects concerning these patterns ([43, 12]) can be generated from video data streams as statistical values and can be used to computationally identify correlations with the user profile, in terms of perceptual and emotional effects that users like.

3 Artistic Background

In this section, we provide the artistic background to the idea of using stylistic visual features for movie recommendation. We describe the stylistic visual features from an artistic point of view and explain the possible relation between these low-level features and the aesthetic variables that are well-known for artists in the domain of movie making.

As noted briefly in Section 2, the study on how various aesthetic variables and their combination contribute to establish the meaning conveyed by an artistic work is the domain of different disciplines, e.g., semiotics, traditional aesthetic studies, etc. The shared belief is that humans respond to certain stimuli (being them called *signs*, *symbols*, *features* depending on the discipline) in ways that are predictable, up to a given extent. One of the consequences about this, is that similar stimuli are expected to provoke similar reactions, and this in turn may allow to group similar works of art together by the reaction they are expected to provoke.

Among these disciplines, of particular interest for this paper, Applied Media Aesthetic [53] is concerned with the relation of a number of media elements, such as light, camera movements, colors, with the perceptual reactions they are able to evoke in consumers of media communication, mainly videos and films. Such media elements, that together build the visual images composing the media, are investigated following a rather

a. b.

Fig. 1 a. *Out of the past* (1947) an example of highly contrasted lighting. **b.** *The wizard of OZ* (1939) flat lighting example.

formalistic approach that suits the purposes of this paper. By an analysis of cameras, lenses, lighting, etc., as production tools as well as their aesthetic characteristics and uses, Applied Media Aesthetic tries to identify patterns in how such elements operate to produce the desired effect in communicating emotions and meanings.

The image elements that are usually addressed as fundamental in the literature, e.g. in [23], even if with slight differences due to the specific context, are lights and shadows, colors, space representation, motion and sound. It has been proved, e.g. in [43][12], that some aspects concerning these elements can be computed from the video data stream as statistical values. We call these computable aspects as *features*.

We will now look into closer details of the features, investigated for content-based video recommendation in this paper to provide a solid overview on how they are used to producing perceptual reaction in the audience. Sound will not be further discussed, since it is out of scope of this work, as well as the space representation, that concerns, e.g., the different shooting angles that can be used to represent dramatically an event.

3.1 Lighting

There are at least two different purposes for lighting in video and movies. The most obvious is to allow and define viewers' perception of the environment, to make visible the objects and places they look at. But light can also manipulate how, emotionally, an event is to be perceived, acting in a way that bypass rational screens. The two main lighting alternatives are usually addressed to as *chiaroscuro* and *flat lighting*, but there are many intermediate solutions between them. The first is a lighting technique characterized by high contrast between light and shadow areas that put the emphasis on an unnatural effect: the borders of objects are altered by the lights. The latter instead is an almost neutral, realistic, way of illuminating, whose purpose is to enable recognition of stage objects. Figure 1a and Figure 1b exemplifies these two alternatives.

3.2 Colors

The expressive quality of colors is closely related to that of lighting, sharing the same ability to set or magnify the feeling derived by a given situation. The problem with colors is the difficulty to isolate their contribution to the overall 'mood' of a scene from that of other aesthetic variables operating in the same context. Usually their effectiveness is higher when the context as a whole is predisposed toward a specific emotional objective.

Even if an exact correlation between colors and the feeling they may evoke is not currently supported by enough scientific data, colors nonetheless have an expressive impact that has been investigated thoroughly, e.g. in [49]. An interesting metric to quantify this impact has been proposed in [50] as perceived *color energy*, a quantity that depends on a color's saturation, brightness and the size of the area the color covers in an image. Also the hue plays a role as if it tends toward reds, the quantity of energy is more, while if it tends more on blues, it is less. These tendencies are shown in the examples of Figure 2a and Figure 2b.

3.3 Motion

The illusion of movement given by screening a sequence of still frames in rapid succession is the very reason of

a. b.

Fig. 2 a. An image from *Django Unchained* (2012). The red hue is used to increase the scene sense of violence. **b.** An image from *Lincoln* (2012). Blue tone is used to produce the sense of coldness and fatigue experienced by the characters.

cinema existence. In a video or movie, there are different types of motions to consider:

- **Profilmic movements**. Every movement that concerns elements, shot by the camera, falls in this category, e.g. performers motion, vehicles, etc.. The movement can be real or perceived. By deciding the type and quantity of motion an 'actor' has, considering as actor any possible protagonist of a scene, the director defines, among others, the level of attention to, or expectations from, the scene. As an example, the hero walking slowly in a dark alley, or a fast car chasing.
- **Camera movements**. Are the movements that alter the point of view on the narrated events. Camera movements, such as the pan, truck, pedestal, dolly, etc. can be used for different purposes. Some uses are descriptive, to introduce landscapes or actors, to follow performers actions, and others concern the narration, to relate two or more different elements, e.g., anticipating a car's route to show an unseen obstacle, to move toward or away from events.
- **Sequences movements**. As shots changes, using cuts or other transitions, the rhythm of the movie changes accordingly. Generally, a faster rhythm is associated with excitement, and a slower rhythm suggests a more relaxed pace [16].

In this paper, we followed the approach in [43], considering the *motion* content of a scene as a feature that aggregate and generalize both profilmic and camera movements. Sequence movements are instead considered in the *Average shot length* feature, both being described with detail in the next section.

4 Method Description

The first step in order to build a video CBRS based on stylistic low-level features is to search for features that comply with human visual norms of perception and abide by the grammar of the film - the rules that movie producers and directors use in order to make movies. In general, a movie $\mathbf{M}$ can be represented as a combination of three main modalities: M_V the visual, M_A the audio and M_T textual modalities respectively. The focus of this work is only on visual features, therefore $\mathbf{M} = \mathbf{M}(M_V)$. The visual modality itself can be represented as

$$M_V = M_V(\mathbf{f_v}) \tag{1}$$

where $\mathbf{f_v} = (f_1, f_2, ..., f_n)$ is a set of features that describe the visual content of a video. Generally speaking, a video can be considered as contiguous sequence of many frames. Consecutive video frames contain a lot of frames that are highly similar and correlated. Considering all these frames for feature extraction not only does not provide new information to the system but also is computationally inefficient. Therefore, the first step prior to feature extraction is structural analysis of the video, i.e. to detect shot boundaries and to extract a key frame within each shot. A shot boundary is a frame where frames around it have significant difference in their visual content. Frames within a shot are highly similar on the other hand, therefore it makes sense to take one representative frame in each shot and use that frame for feature extraction. This frame is called the *Key Frame*.

Fig. 3 illustrates the hierarchical representation of a video. Two types of features are extracted from videos: (i) temporal features (ii) spatial features. The temporal features reflect the dynamic perspectives in a video

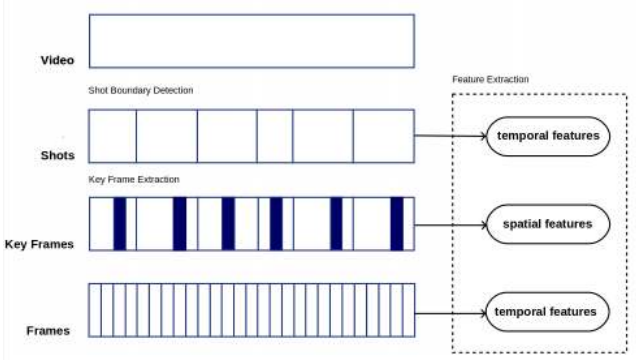

Fig. 3 Hierarchical video representation and feature extraction in our framework

such as the average shot duration (or shot length) and object motion, whereas the spatial features illustrate static properties such as color, light, *etc.*. In the following we describe in more detail these features and the rationale behind choosing them in addition to how they can be measured in a video.

4.1 Visual features

In order to demonstrate the effectiveness of the proposed video CBRS, after carefully studying the literature in computer vision, we selected and extracted the five most informative and distinctive features to be extracted from each video

$$\begin{aligned}\mathbf{f_v} &= (f_1, f_2, f_3, f_4, f_5) \\ &= (\overline{L}_{sh}, \mu_{cv}, \mu_{\overline{m}}, \mu_{\sigma^2_m}, \mu_{lk}) \end{aligned} \tag{2}$$

where $\overline{L}_{sh}$ is the average shot length, μ_{cv} is the mean color variance over key frames, $\mu_{\overline{m}}$ and $\mu_{\sigma^2_m}$ are the mean motion average and standard deviation across all frames respectively and $\mu_{\sigma^2_m}$ is the mean lightening key over key frames. As can be noted, some of the features are calculated across key frames and the others across all video frames (see Fig 3). Each of these features carry a meaning and are used in the hands of able directors to convey emotions when shooting movies. Assuming that there exists n_f frames in the video, t being the index of each single frame and n_{sh} key frames (or shots), q being the index of a numbered list of key frames, the proposed visual features and how they are calculated is presented in the following [20,43,21]

- *Average shot length ($\overline{L}_{sh}$)*: A shot is a single camera action and the number of shots in a video can provide useful information about the pace at which a movie is being created. The average shot length is defined as

$$\overline{L}_{sh} = \frac{n_f}{n_{sh}} \tag{3}$$

where n_f is the number of frames and n_{sh} the number of shots in a movie. For example, action movies usually contain rapid movements of the camera (therefore they contain higher number of shots or shorter shot lengths) compared to dramas which often contain conversations between people (thus longer average shot length). Because movies can be made a different frame rates, $\overline{L}_{sh}$ is further normalized by the frame rate of the movie.

- *Color variance (μ_{cv})*: The variance of color has a strong correlation with the genre. For instance, directors tend to use a large variety of bright colors for comedies and darker hues for horror films. For each key frame represented in Luv color space we compute the covariance matrix:

$$\rho = \begin{pmatrix} \sigma^2_L & \sigma^2_{Lu} & \sigma^2_{Lv} \\ \sigma^2_{Lu} & \sigma^2_u & \sigma^2_{uv} \\ \sigma^2_{Lv} & \sigma^2_{uv} & \sigma^2_v \end{pmatrix} \tag{4}$$

The generalized variance can be used as the representative of the color variance in each key frame given by

$$\Sigma_q = det(\rho) \tag{5}$$

in which a key frame is a representative frame within a shot (e.g. the middle shot). The average color variance is then calculated by:

$$\mu_{cv} = \frac{\sum_{q=1}^{n_{sh}} \Sigma_q}{n_{sh}} \tag{6}$$

where n_{sh} is the number of shots equal to number of key frames.

- *Motion*: Motion within a video can be caused mainly by the camera movement (*i.e.* camera motion) or movements on part of the object being filmed (*i.e.* object motion). While the average shot length captures the former characteristic of a movie, it is desired for the motion feature to also capture the latter characteristic. A motion feature descriptor based on optical flow [6,28] is used to measure a robust estimate of the motion in sequence of images based on velocities of images being filmed. Because motion features are based upon sequence of images, they are calculated across all video frames.
  At frame t, if the average motion of pixels is represented by $\overline{m}_t$ and the standard deviation of pixel motions is $(\sigma^2_m)_t$:

$$\mu_{\overline{m}} = \frac{\sum_{t=1}^{n_f} \overline{m}_t}{n_f} \tag{7}$$

and

$$\mu_{\sigma_m^2} = \frac{\sum_{t=1}^{n_f} (\sigma_m^2)_t}{n_f} \tag{8}$$

where $\mu_{\overline{m}}$ and $\mu_{\sigma_m^2}$ represent the average of motion mean and motion standard deviation aggregated over entire n_f frames.

– *Lighting Key*: Lighting key is another distinguishing factor between movie genres in such a way that the director use it as a factor to control the type of emotion they want to be induced to a viewer. For example, comedy movies often adopt lighting key which has abundance of light (*i.e.* high gray-scale mean) with less contrast between the brightest and dimmest light (*i.e.* high gray-scale standard deviation). This trend is often known as *high-key* lighting. On the other hand, horror movies or noir films often pick gray-scale distributions which is low in both gray-scale mean and gray-scale standard deviation, known by *low-key* lighting. In order to capture both of these parameters, after transforming all key-frames to HSV color-space [48], we compute the mean μ and standard deviation σ of the value component which corresponds to the brightness. The scene lighting key ξ defined by multiplication of μ and σ is used to measure the lighting of key frames

$$\xi_q = \mu \cdot \sigma \tag{9}$$

For instance, comedies often contain key-frames which have a well distributed gray-scale distribution which results in both the mean and standard deviation of gray-scale values to be high therefore for comedy genre one can state $\xi > \tau_c$, whereas for horror movies the lighting key with poorly distributed lighting the situation is reverse and we will have $\xi < \tau_h$, where τ_c and τ_h are predefined thresholds. In the situation where $\tau_h < \xi < \tau_c$ other movie genres (*e.g.* Drama) exists where it is hard to use the above distinguish factor for them. The average lighting calculated over key frames is given by (10)

$$\mu_{lk} = \frac{\sum_{q=1}^{n_{sh}} \xi_q}{n_{sh}} \tag{10}$$

It worth noting that the stylistic visual features have been extracted by using our own implementation. The code and the dataset of extracted features will be publicly accessible through the webpage of the group [1].

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

4.2 Recommendation algorithm

To generate recommendations using our Low-Level stylistic visual features, we adopted a classical "k-nearest neighbor" content-based algorithm. Given a set of users $u \in U$ and a catalogue of items $i \in I$, a set of preference scores r_{ui} given by user u to item i has been collected. Moreover, each item $i \in I$ is associated to its feature vector $\boldsymbol{f_i}$. For each couple of items i and j, the similarity score s_{ij} is computed using *cosine similarity*:

$$s_{ij} = \frac{\boldsymbol{f_i}^T \boldsymbol{f_j}}{\|\boldsymbol{f_i}\| \, \|\boldsymbol{f_j}\|} \tag{11}$$

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

$$\hat{r}_{ui} = \frac{\sum_{j \in NN_i, r_{uj} > 0} r_{uj} s_{ij}}{\sum_{j \in NN_i, r_{uj} > 0} s_{ij}} \tag{12}$$

5 Evaluation Methodology

We have formulated the following two hypotheses:

1. the content-based recommender system, that exploits a set of representative visual features of the video contents, may have led to a higher recommendation accuracy in comparison to the genre based recommender system.
2. the trailers of the movies can be representative of their full-length movies, with respect to the stylistic visual features, and indicate high correlation with them.

Hence, we speculate that a set of stylistic visual features, extracted automatically, may be more informative of the video content than a set of high-level expert annotated features.

In order to test these hypotheses, we have evaluated the Top-N recommendation quality of each content-based recommender systems by running a 5-fold cross-validation on a subset of the MovieLens-20M dataset [1]. The details of the subset are described later in the paper. The details on the evaluation procedure follow. First, we generated 5 disjoint random splits of the ratings in the dataset. Within each iteration, the evaluation procedure closely resambles the one described in [17]. For each iteration, one split was used as the probe set P_i, while the remaining ones were used to generate the training set M_i and was used to train the recommendation algorithm.The test set T_i contains only 4-star and 5-star ratings from P_i, which we assume to be

relevant.
For each relevant item i rated by user u in T_i, we form a list containing the item i and all the items not rated by the user u, which we assume to be irrelevant to her. Then, we formed a top-N recommendation list by picking the top N ranked items from the list. Being r the rank of i, we have a *hit* if $r < N$, otherwise we have a *miss*. Since we have one single relevant item per test case, recall can assume value 0 or 1 (respectively in case of a miss or a hit). Therefore, the *recall(N)* on the test set T_i can be easily computed as:

$$recall(N) = \frac{\#hits}{|T_i|} \tag{13}$$

We could have also evaluated the *precision(N)* on the recommendation list, but since it is related to the value of the *recall(N)* by a simple scaling factor $1/N$ [17], we decided to omit it to avoid redundancy. The values reported throughout the paper are the averages over the 5 folds.

We have used a set of full-length movies and their trailers, that were sampled randomly from all the main genres, i.e., Action, Comedy, Drama and Horror. The summary of the dataset is given in Table 1.

Table 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* [2]. The dataset contained over all 167 movies, 105 of which belonging to a single genre and 62 movies belonging to multiple genres (see Table 2).

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

6 Results

6.1 Classification Accuracy

We have conducted a preliminary experiment to understand if the genre of the movies can be explained in terms of the five low-level visual features, described in Section 4. The goal of the experiment is to classify the movies into genres by exploiting their visual features.

6.1.1 Experiment A

In order to simplify the experiment, we have considered 105 movies tagged with one genre only (Table 2). We have experimented with many classification algorithms and obtained the best results under *decision tables* [32]. Decision tables can be considered as tabular knowledge representations [33]. Using this technique, the classification for a new instance is done by searching for the exact matches in the decision table cells, and then the instance is assigned to the most frequent class among all instances matching that table cell [26].

We have used 10 fold cross-validation and obtained an accuracy of 76.2% for trailers and 70.5 % for full-length movies. The best classification was done for comedy genre: 23 out of 27 movie trailers were successfully classified in their corresponding comedy genre. On the other hand, the most erroneous classification happened for the horror genre where more number of movie trailers have been misclassified into the other genres. For example, 4 out of 24 horror movie trailers have been mistakenly classified as action genre. This is a phenomenon that was expected, since typically there are many action scenes occurred in horror movies, and this may make the classification very hard. Similar results have been observed for full-length movies.

From this first experiment we can conclude that the five low-level stylistic visual features used in our experiment can be informative of the movie content and can be used to accurately classify the movies into their corresponding genres.

[2] http://www.mathworks.com/products/matlab
[3] https://www.r-project.org
[4] http://www.cs.waikato.ac.nz/ml/weka

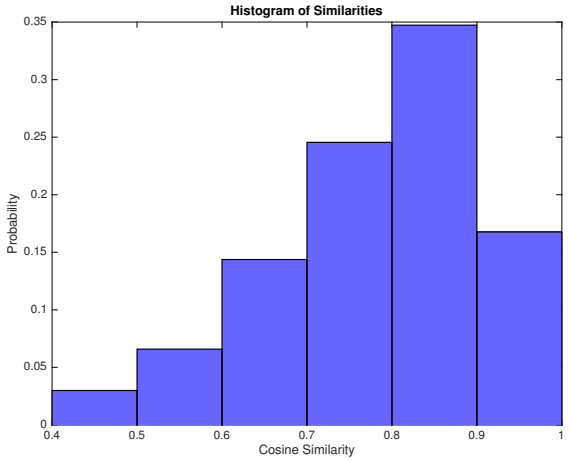

Fig. 4 Histogram distribution of the cosine similarity between full-length movies and trailers

6.2 Correlation between Full-length Movies and Trailers

One of the research hypotheses we have formulated addresses the possible correlation between the full-length movies and their corresponding trailers. Indeed, we are interested to investigate whether or not the trailers are representative of their full-length movies, with respect to the stylistic visual features. In order to investigate this issue, we have performed two experiments.

6.2.1 Experiment B

We have first extracted the low-level visual features from each of the 167 movies and their corresponding trailers in our dataset. We have then computed the cosine similarity between the visual features extracted from the full-length movies and the trailers. Cosine is the same metric used to generate recommendations in our experiments (as explained in Section 5), 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 4 plots the histogram of the cosine similarity. Average is 0.78, median is 0.80. More than 75% of the movies have a cosine similarity greater than 0.7 between the full-length movie and trailer. Moreover, less than 3% of the movies have a similarity below 0.5.

Overall, the cosine similarity shows a substantial correlation between the full-length movies and trailers. 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.

6.2.2 Experiment C

In the second experiment, we have used the low-level features extracted from both trailers and the full-length movies to feed the content-based recommender system described in Section 4.2. We have used features f3-f5 (i.e., camera motion, object motion, and light) as they proved to be the best choice of stylistic visual features (as described in Section 6.3). Quality of recommendations has been evaluated according to the methodology described in Section 5.

Figure 5 plots the recall@N for full-length movies (a) and trailers (b), with values of N ranging from 1 to 5. We note that the K values have been determined with cross validation (see Section 6.3.1). By comparing the two figures, it is clear that the recall values of the content-based recommendation using the features extracted from the full-length movies and trailers are almost identical.

The results of this second experiment confirm that low-level features extracted from trailers are representative of the corresponding full-length movies and can be effectively used to provide recommendations.

6.3 Recommendation Quality

In this section we investigate our main research hypothesis: if low-level visual features can be used to provide good-quality recommendations. We compare the quality of content-based recommendations based on three different types of features:

Low Level (LL): *stylistic visual* features.
High Level (HL): *semantic* features based on *genres*.
Hybrid (LL+HL): both stylistic and semantic features.

6.3.1 Experiment D

In order to identify the visual features that are more useful in terms of recommendation quality, we have performed an exhaustive set of experiments by feeding a content-based recommender system with all the 31 combinations of the five visual features f1-f5. Features have been extracted from the trailers. We have also combined the low-level stylistic visual features with the genre, resulting in 31 additional combinations. When using two or more low-level features, each feature has been normalized with respect to its maximum value (infinite norm).

Table 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).

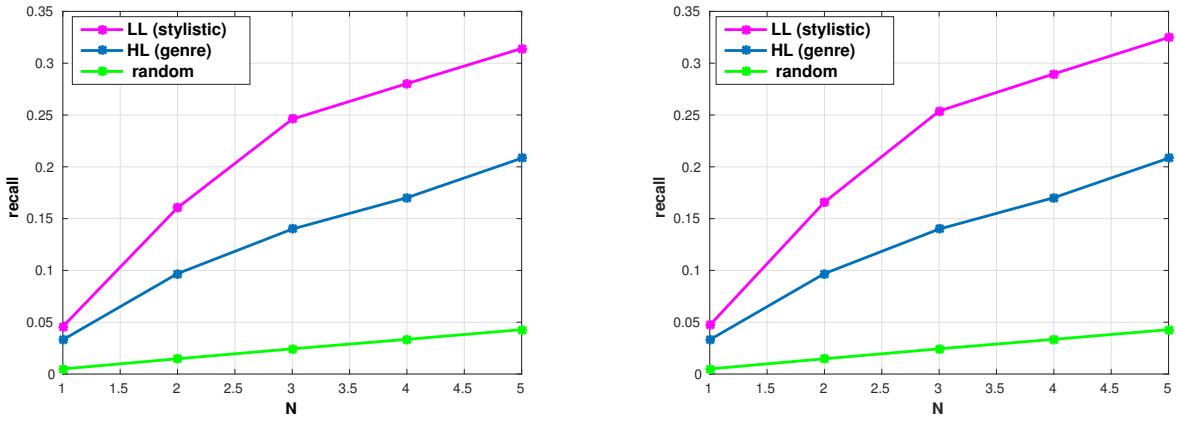

Fig. 5 Performance comparison of different CB methods under best feature combination for full-length movies (a), and trailers (b). $K = 2$ for LL features and $K = 10$ for HL features.

The last column of the table reports, as a reference, the recall when using genre only. This value does not depend on the low-level features. The optimal value of K for the KNN similarity has been determined with cross validation: $K = 2$ for low-level and hybrid, $K = 10$ for genre-only.

Recommendations based on low-level stylistic visual features extracted from trailers are clearly better, in terms of recall, than recommendations based on genre for **any** combination of visual features. However, no considerable difference has been observed between genre based and hybrid based recommendations.

It is worth noting that, when using low-level feature f2 (color variance), recommendations have a lower accuracy with respect to the other low-level features, although always better with respect to genre-based recommendation. Moreover, when using two or more low-level features together, accuracy does not increase. These results will be further investigated in the next section.

6.4 Feature Analysis

In this section, we wish to investigate why some low-level visual features provide better recommendations than the others, as highlighted in the previous section. Moreover, we investigate why combinations of low-level features do not improve accuracy.

6.4.1 Experiment E

In a first experiment, we analyze if some of the low-level features extracted from trailers are better correlated than the others with respect to the corresponding features extracted from the full-length movie. This analysis is similar to the one reported in Section 6.2, but results are reported as a function of the features.

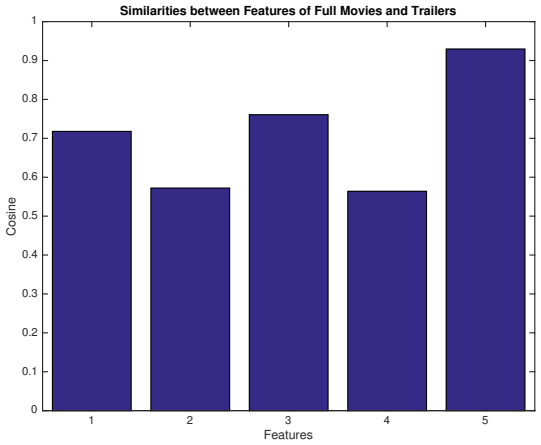

Fig. 7 Cosine similarity between stylistic visual features extracted from the full-length movies and their corresponding trailers

Figure 7 plots the cosine similarity values between visual features extracted from the full-length movies and visual features extracted from trailers. Features f2 and f4 (color variance and object motion) are the less similar features, suggesting that their adoption, if extracted from trailers, should provide less accurate recommendations.

We also performed Wilcoxon test comparing features extracted from the full-length movies and trailers. The results, summarized in Table 4, prove that no significant difference exists between features f3 (camera motion) and f5 (light), which clearly shows that the full-length movies and trailers are highly correlation with respect to these two features. For the other features,

Table 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	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	0.21
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	

significant differences have been obtained. This basically states that some of the extracted features may be either not correlated or not very informative.

6.4.2 Experiment F

In figure 6, scatter plots of all combinations of the 5 stylistic visual features (intra-set similarity) are plotted. Having visually inspected, it can be seen that, overall the features are weakly correlated. However, there are still features that mutually present high degree of linear correlation. For example, feature 3 and 4, seem to be highly correlated (see row 3, column 4 in figure 6). Moreover, we have observed similarity by comparing the scatter plots of full-length movies and trailers. Indeed, any mutual dependency between different features extracted either from full-length movies or trailers were similar. This is another indication that trailers can be considered as representative short version of the full-length movies, in terms of stylistic visual features.

6.4.3 Informativeness of the Features

Entropy is an information theoretic measure [27] that is an indication of the informativeness of the data. Figure 8 illustrates the entropy scores computed for the stylistic visual features. As it can be seen, the entropy scores of almost all visual stylistic features are high. The most informative feature, in terms of entropy score, is the fifth one, i.e. Lighting Key, and the least informative feature is the second one, i.e., Color Variance (see Section 4 for detailed description). This observation is in the full consistency with the other findings, that we have obtained from, e.g. Wilcoxon test (see Table 4), and correlation analysis (see Figure 7)

Having considered all the results, we remark that our considered hypotheses have been successfully validated, and we have shown that a proper extraction of the visual stylistic features of videos may have led to higher accuracy of video recommendation, than the typical expert annotation method, either when the features are extracted from full-length videos or when the features originate from movie trailers only. These are

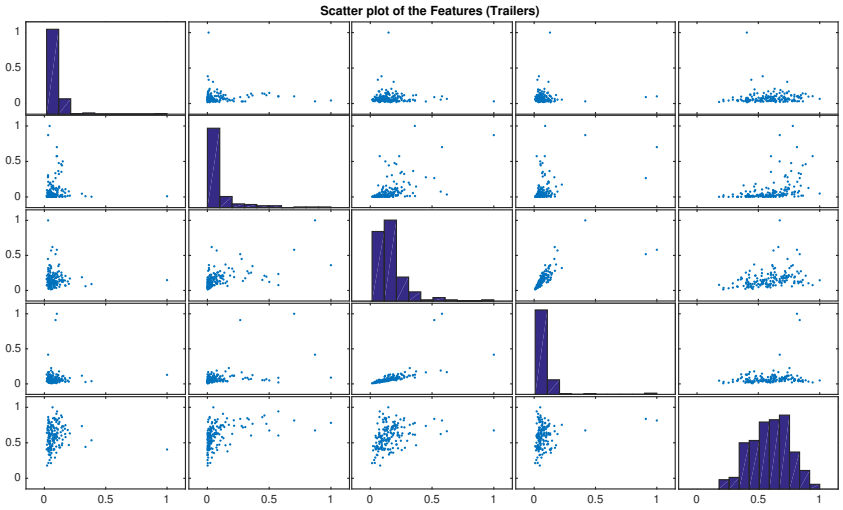

Fig. 6 Scatter plot for different combination of the stylistic visual features extracted from the movie trailers

Table 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 $(\mu_{\sigma^2_{\overline{m}}})$	$f_5(\mu_{lk})$
wilcox.test	1.3e-9	5.2e-5	**0.154**	2.2e-16	**0.218**
H0/H1	w ->H1	w ->H1	w ->H0	w ->H1	w ->H0

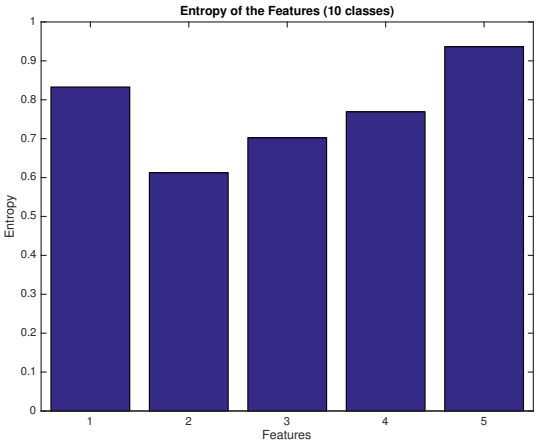

Fig. 8 Entropy of the stylistic visual features

promising results, as they overall illustrate the possibility to achieve higher accuracy with an automatic method than a manual method (i.e., expert annotation of videos) since the manual method can be very costly and in some cases even impossible (e.g., in huge datasets).

7 Discussion

The results presented in the previous section confirm both our research hypothesis:

1. recommendations based on low-level stylistic visual features are better than recommendations based on high level semantic features, such as genre;
2. low-level features extracted from trailers are, in general, a good approximation of the corresponding features extracted from the original full-length movies.

7.1 Quality of Recommendations

According to Table 3, all the low-level visual features provide better recommendations than the high-level features (genres). 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 are not strongly affected by how colors are used in movies. This is partially explained by the limited informative content of the color variance feature, as show in Figure 8, where color variance is the feature with the lowest entropy.

The validity of this finding is restricted to the actual experimental conditions considered, and may be affected by the limited size of the dataset. In spite of these limitations, our results provide empirical evidence that

> *the tested low-level visual features may 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. 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, a correlation exists between low-level and high level features that leads to collinearities and reduced prediction capabilities of the mixed approach.

When using a combination of two or more low-level features, the quality of recommendations does not increase significantly and, in same cases, decreases, although it is always better than the quality obtained with high-level features. This behavior is not surprising, considering that the low-level features are weakly correlated, as shown in Figure 6.

7.2 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.

Our research shows 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). 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. However, the strong correlation 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

> *low-level visual features extracted from trailers can be used as an alternative to features extracted from full-length movies in building content-based recommender systems.*

8 Conclusion and Future Work

In this paper, we have presented a novel content-based 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 high-level semantic features – such as, genre, cast, or reviews – that 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.

Our future work can be extended in a number of challenging directions:

- We will widen our analysis by adopting bigger and different datasets, in order to provide a more robust statistical support to our finding.
- We will investigate the impact of using different content-based recommendation algorithms, such as those based on Latent-Semantic-Analysis, when adopting low-level features.
- We will extend the range of visual features extracted and we will also include audio features.
- we will analyze recommender systems based on low-level features not only in terms of accuracy, but also in terms of perceived novelty and diversity, with a set of online user studies.

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

References

1. Datasets — grouplens. http://grouplens.org/datasets/. Accessed: 2015-05-01
2. Youtube. http://www.youtube.com. Accessed: 2015-04-01

3. Adomavicius, G., Tuzhilin, A.: 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)
4. Ahn, J.w., Brusilovsky, P., Grady, J., He, D., Syn, S.Y.: Open user profiles for adaptive news systems: help or harm? In: Proceedings of the 16th international conference on World Wide Web, pp. 11–20. ACM (2007)
5. Balabanović, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Commun. ACM **40**(3), 66–72 (1997)
6. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International journal of computer vision **12**(1), 43–77 (1994)
7. Billsus, D., Pazzani, M.J.: A hybrid user model for news story classification. Springer (1999)
8. Billsus, D., Pazzani, M.J.: User modeling for adaptive news access. User modeling and user-adapted interaction **10**(2-3), 147–180 (2000)
9. Bogdanov, D., Herrera, P.: How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In: ISMIR, pp. 97–102 (2011)
10. Bogdanov, D., Serrà, J., Wack, N., Herrera, P., Serra, X.: Unifying low-level and high-level music similarity measures. Multimedia, IEEE Transactions on **13**(4), 687–701 (2011)
11. Brezeale, D., Cook, D.J.: Automatic video classification: A survey of the literature. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **38**(3), 416–430 (2008)
12. Buckland, W.: 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 (2008)
13. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction **12**(4), 331–370 (2002). URL ./papers/burke-umuai-ip-2002.pdf
14. Canini, L., Benini, S., Leonardi, R.: Affective recommendation of movies based on selected connotative features. Circuits and Systems for Video Technology, IEEE Transactions on **23**(4), 636–647 (2013)
15. Cantador, I., Szomszor, M., Alani, H., Fernández, M., Castells, P.: Enriching ontological user profiles with tagging history for multi-domain recommendations (2008)
16. Choroś, K.: Video shot selection and content-based scene detection for automatic classification of tv sports news. In: E. Tkacz, A. Kapczynski (eds.) Internet Technical Development and Applications, *Advances in Intelligent and Soft Computing*, vol. 64, pp. 73–80. Springer Berlin Heidelberg (2009)
17. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010, pp. 39–46 (2010)
18. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al.: The youtube video recommendation system. In: Proceedings of the fourth ACM conference on Recommender systems, pp. 293–296. ACM (2010)
19. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction **17**(3), 217–255 (2007)
20. Deldjoo, Y., Elahi, M., Quadrana, M., Cremonesi, P.: Toward building a content-based video recommendation system based on low-level features. In: E-Commerce and Web Technologies. Springer (2015)
21. Deldjoo, Y., Elahi, M., Quadrana, M., Cremonesi, P., Garzotto, F.: Toward effective movie recommendations based on mise-en-scène film styles. In: Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter, pp. 162–165. ACM (2015)
22. Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Transactions on Information Systems (TOIS) **22**(1), 143–177 (2004)
23. Dorai, C., Venkatesh, S.: Computational media aesthetics: Finding meaning beautiful. IEEE MultiMedia **8**(4), 10–12 (2001)
24. Eirinaki, M., Vazirgiannis, M., Varlamis, I.: Sewep: using site semantics and a taxonomy to enhance the web personalization process. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 99–108. ACM (2003)
25. Elahi, M., Ricci, F., Rubens, N.: Active learning strategies for rating elicitation in collaborative filtering: a system-wide perspective. ACM Transactions on Intelligent Systems and Technology (TIST) **5**(1), 13 (2013)
26. Freitas, A.A.: Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter **15**(1), 1–10 (2014)
27. Guyon, I., Matic, N., Vapnik, V., et al.: Discovering informative patterns and data cleaning. (1996)
28. Horn, B.K., Schunck, B.G.: Determining optical flow. In: 1981 Technical Symposium East, pp. 319–331. International Society for Optics and Photonics (1981)
29. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: 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)
30. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. In: ACM SIGIR Forum, vol. 37, pp. 18–28. ACM (2003)
31. Knees, P., Pohle, T., Schedl, M., Widmer, G.: A music search engine built upon audio-based and web-based similarity measures. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 447–454. ACM (2007)
32. Kohavi, R.: The power of decision tables. In: 8th European Conference on Machine Learning, pp. 174–189. Springer (1995)
33. Kohavi, R., Sommerfield, D.: Targeting business users with decision table classifiers. In: KDD, pp. 249–253 (1998)
34. Lehinevych, T., Kokkinis-Ntrenis, N., Siantikos, G., Dogruöz, A.S., Giannakopoulos, T., Konstantopoulos, S.: Discovering similarities for content-based recommendation and browsing in multimedia collections. In: Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on, pp. 237–243. IEEE (2014)
35. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **2**(1), 1–19 (2006)
36. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender systems handbook, pp. 73–105. Springer (2011)
37. Magnini, B., Strapparava, C.: Improving user modelling with content-based techniques. In: User Modeling 2001, pp. 74–83. Springer (2001)

38. Middleton, S.E., Shadbolt, N.R., De Roure, D.C.: Ontological user profiling in recommender systems. ACM Transactions on Information Systems (TOIS) **22**(1), 54–88 (2004)
39. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the fifth ACM conference on Digital libraries, pp. 195–204. ACM (2000)
40. Musto, C., Narducci, F., Lops, P., Semeraro, G., de Gemmis, M., Barbieri, M., Korst, J., Pronk, V., Clout, R.: Enhanced semantic tv-show representation for personalized electronic program guides. In: User Modeling, Adaptation, and Personalization, pp. 188–199. Springer (2012)
41. Pazzani, M.J., Billsus, D.: The adaptive web. chap. Content-based Recommendation Systems, pp. 325–341. Springer-Verlag, Berlin, Heidelberg (2007). URL http://dl.acm.org/citation.cfm?id=1768197.1768209
42. Rasheed, Z., Shah, M.: Video categorization using semantics and semiotics. In: Video mining, pp. 185–217. Springer (2003)
43. Rasheed, Z., Sheikh, Y., Shah, M.: On the use of computable features for film classification. Circuits and Systems for Video Technology, IEEE Transactions on **15**(1), 52–64 (2005)
44. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (eds.) Recommender Systems Handbook, pp. 1–35. Springer Verlag (2011)
45. Rubens, N., Elahi, M., Sugiyama, M., Kaplan, D.: Active learning in recommender systems. In: Recommender Systems Handbook, pp. 809–846. Springer (2015)
46. Seyerlehner, K., Schedl, M., Pohle, T., Knees, P.: 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)
47. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. in Artif. Intell. **2009**, 4:2–4:2 (2009)
48. Tkalcic, M., Tasic, J.F.: Colour spaces: perceptual, historical and applicational background. In: EUROCON 2003. Computer as a Tool. The IEEE Region 8, vol. 1, pp. 304–308. IEEE (2003)
49. Valdez, P., Mehrabian, A.: Effects of color on emotions. Journal of Experimental Psychology: General **123**(4), 394 (1994)
50. Wang, H.L., Cheong, L.F.: Affective understanding in film. Circuits and Systems for Video Technology, IEEE Transactions on **16**(6), 689–704 (2006)
51. Wang, Y., Xing, C., Zhou, L.: Video semantic models: survey and evaluation. Int. J. Comput. Sci. Netw. Security **6**, 10–20 (2006)
52. Yang, B., Mei, T., Hua, X.S., Yang, L., Yang, S.Q., Li, M.: Online video recommendation based on multimodal fusion and relevance feedback. In: Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 73–80. ACM (2007)
53. Zettl, H.: Essentials of applied media aesthetics. In: C. Dorai, S. Venkatesh (eds.) Media Computing, *The Springer International Series in Video Computing*, vol. 4, pp. 11–38. Springer US (2002)
54. Zhao, X., Li, G., Wang, M., Yuan, J., Zha, Z.J., Li, Z., Chua, T.S.: Integrating rich information for video recommendation with multi-task rank aggregation. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 1521–1524. ACM (2011)
55. Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M.: Movie genre classification via scene categorization. In: Proceedings of the international conference on Multimedia, pp. 747–750. ACM (2010)