Addressing the New Item problem in video recommender systems by incorporation of visual features with restricted Boltzmann machines

Article *in* Expert Systems · May 2021

DOI: 10.1111/exsy.12645

CITATIONS
7

READS
318

2 authors:

Naieme Hazrati
Free University of Bozen-Bolzano
6 PUBLICATIONS **22** CITATIONS

SEE PROFILE

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

SEE PROFILE

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Addressing the New Item Problem in Video Recommender Systems by Incorporation of Visual Features with Restricted Boltzmann Machines

Naieme Hazrati[1] | Mehdi Elahi[2]

[1]Department of Computer Science, Free University of Bozen - Bolzano, Bolzano, Italy
[2]Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

Correspondence
*Mehdi Elahi, Email: mehdi.elahi@uib.no

Summary

Over the past years, the research of video Recommender Systems (RSs) has been mainly focused on the development of novel algorithms. Although beneficial, still any algorithm may fail to recommend video items that the system has no form of data associated to them (*New Item Cold Start*). This problem occurs when a new item is added to the catalog of the system and no data is available for that item. In *Content-based* RSs, the video items are typically represented by semantic attributes, when generating recommendations. These attributes require a group of experts or users for annotation, and still, the generated recommendations might not capture a complete picture of the users' preferences, e.g., the visual tastes of users on video style. This article addresses this problem by proposing recommendation based on novel visual features that do not require human-annotation and can represent visual aspects of video items. We have designed a novel evaluation methodology considering three realistic scenarios, i.e., (i) *Extreme cold start*, (ii) *Moderate cold start*, and (iii) *Warm-start* scenario. We have conducted a set of comprehensive experiments and our results have shown the superior performance of recommendations based on visual features, in all of the evaluation scenarios.

KEYWORDS:
Recommender Systems, Cold Start, Visually-aware, New Item, Multimedia

1 | INTRODUCTION

Nowadays, finding the right videos to consume has become a big challenge for users due to the enormous *Volume*, *Variety*, and *Velocity* (i.e., 3Vs of Big Data of producing and sharing video content online. It has been reported that *YouTube*, as an example of popular video-sharing applications, has about 1.5 billion active users who consume incredible number of 5 billion videos per day . Due to the 3Vs of this big data, users may feel desperate when they need to choose among an unlimited choices and they may fail to find their desired videos with relevant content. On the other hand, due to this characteristics of input data, it could be extremely difficult for video-sharing applications to remedy this problem and support the users in finding desired video content. Hence, it is not uncommon to observe confused video consumers unable to find interesting content from the enormous volume and variety of videos they can choose from Anderson (2006).

Recommender Systems can cope with this big challenge by supporting the users when making decision on what to consume Adomavicius and Tuzhilin (2005); Aggarwal (2016b); Jannach, Zanker, Felfernig, and Friedrich (2010); Lops, De Gemmis, and Semeraro (2011); Resnick and Varian (1997a); Ricci, Rokach, and Shapira (2015); Santos and Boticario (2015); Schafer, Konstan, and Riedl (2001). These decision support systems can build personalized video suggestions based on the *specific* tastes and interests of users for videos that can better match users' needs and constraints rather than suggesting the popular video content based on *generic* mainstream tastes Dwivedi and Bharadwaj (2015); Jannach et al. (2010); Resnick and Varian (1997b); Ricci, Rokach, Shapira, and Kantor (2011).

In recent years, various types of video recommendation algorithms have been proposed and evaluated presenting excellency in performance. These algorithms typically receive different types of input data, e.g., content-associated data (genre, and tags), and build recommendations on top of this data Aggarwal (2016a); De Gemmis, Lops, Semeraro, and Basile (2008); Gedikli and Jannach (2013); Lops et al. (2011); Lops, De Gemmis, Semeraro, Musto, and Narducci (2013); Shepitsen, Gemmell, Mobasher, and Burke (2008). For that, the data is first cleaned, engineered and then used to create a *Vector Space Model* where the video items represented as vectors of attributes. Then a set of recommendations for a target user is generated by finding video items that share similar attributes with the other items that user has preferred in the past.

While the performance of these recommender algorithms can impact the quality of the generated recommendations, however, any type of algorithms may fail to generate relevant recommendations of video items which have no or very limited amount of associated data Elahi, Ricci, and Rubens (2016); Li, Cheng, Su, and Sun (2017); Lika, Kolomvatsos, and Hadjiefthymiades (2014); Rubens, Elahi, Sugiyama, and Kaplan (2015); Vlachos et al. (2018). This is a situation known as *New Item Cold Start* problem, which typically occurs when a new item is added to the catalog of the system and no input data is available for that item Hornick and Tamayo (2012). This is a major problem in video-sharing applications, such as YouTube where hundreds of hours of videos are uploaded in every minute, by millions of active video makers.

Furthermore, collecting the traditional types of content-associated data, that are typically represented by semantic attributes (e.g., genre, tags), requires either a group of experts or a network of users Cantador, Bellogín, and Vallet (2010); Cantador, Konstas, and Jose (2011); Di Noia, Mirizzi, Ostuni, Romito, and Zanker (2012); Milicevic, Nanopoulos, and Ivanovic (2010); L. Wang, Zeng, Koehl, and Chen (2015). This indeed is an expensive process and needs human efforts. Then recommendations based on these costly semantic attributes still may not properly capture the true users' preferences, e.g., the user tastes associated with visual style of videos.

In addressing this problem, this article investigates the potential behind different types of visual features representative of video content in building quality recommendations for users. We have exploited two different types of visual features, i.e., (i) *Mise-en-scène* features Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016), and (ii) *MPEG7* features Manjunath, Ohm, Vasudevan, and Yamada (2001). Both of these types of visual features can be extracted completely *automatic* without any need for costly *manual* human annotation. Hence they can be exploited by any content-based recommender algorithm capable of incorporating them in the recommendation process Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016); Deldjoo, Quadrana, Elahi, and Cremonesi (2017). While these visual features have been used in prior works, they have never been compared against each-other. In this article, we extend these works by providing an extensive comparison of the recommendations based on these features, as well as traditional semantic attributes. These novel features are compared not only when used *individually* but also when used in *combination* with traditional semantic attributes.

Prior studies have shown that these visual features are not equally informative of the video content Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016). Indeed, some of these features better represent the videos and hence can contribute more when used by a recommender algorithm. Co-linearities and correlations of these features may also negatively affect the recommendation quality. Hence, we designed and developed an artificial neural network based on *Restricted Boltzmann Machines (RBM)* and trained it on the extracted visual features. This network can exploit complex connections among neurons and consequently analyze the input raw features. It can then learn and assign different weights to different input features, according to their representativeness of video content. This type of feature engineering is conducted automatically leading to the computation of the output layer. Finally, the recommender algorithm is trained on this output layer of the neural network and generated personalized recommendations for users.

In conducting this research work, we have formulated the following research questions:

- **[RQ1]** Can recommendation based on the visual features, extracted automatically, remedy the *New Item* problem in *extreme cold start* scenario;
- **[RQ2]** Can recommendation based on the automatic visual features effectively remedy the *New Item* problem in *moderate cold start* scenario;
- **[RQ3]** How combining visual features with traditional semantic attributes can improve the quality of recommendation in *warm start* scenario.

In order to address the research questions, we have designed and developed a novel evaluation methodology that is meant to test the performance of content-based recommender systems in 3 different scenarios, that can occur in real-world applications:

- **Extreme cold start**: happens when absolutely no (human-annotated) semantic attributes are available for a video item (as known as *extreme new item* problem);
- **Moderate cold start**: happens when only a limited amount of (human-annotated) semantic attributes are available for a video item (as known as *moderate new item* problem);
- **Warm start**: happens when considerable amount of (human-annotated) semantic attributes are available for a video item.

This evaluation methodology allowed us to test and compare the recommendations based on proposed visual features with the baseline semantic attributes, in realistic evaluation scenarios. These set of comparisons includes exploiting novel visual features for recommendation generation

individually or in combination with other types of traditional semantic attributes. The comparisons have been conducted with respect to various evaluation metrics (i.e., NRMSE, Precision, and Diversity) using a large dataset of more than ≈8'900'000 ratings obtained from a large network of ≈242'000 users rating ≈4000 movie trailers.

The overall results of the evaluation have shown the consistent superiority of the recommendations based on novel visual features over the traditional semantic attributes, in all of the noted scenarios.

The main contributions of this work are listed on the following:

- we propose a novel technique for video recommendation based on visual features, i.e., (i) **Mise-en-scène** and (ii) **MPEG7** features; these novel features can be extracted automatically, without any need for costly human-annotation; the extraction is performed by deeply analyzing the video items, frame-by-frame, and capturing visual aspects of the videos, encoded within these frames;
- we have developed a novel evaluation methodology specifically designed to test the performance of video recommender systems when encountering *New Item* problem in three different scenarios, i.e., (i) *extreme* cold start, (ii) *moderate* cold start, and (iii) warm start;
- we have adopted two different types of content-based recommender algorithms, i.e., (i) **Pure CBF** (Content-Based Filtering) algorithm and (ii) **Hybrid FM** (Factorization Machines) algorithm, in order to better investigate the effectiveness of the generated recommendations based on visual features;
- we have evaluated the recommendations based on novel visual features in these evaluation scenarios, when features are used *individually* or when used in *combination* with traditional semantic attributes; we tested the recommendation quality exploiting a large dataset of thousands of movies trailers, with millions of ratings given by hundreds of thousands of users.

The rest of the article is structured as the following. In section 2 we survey the related work on video recommender systems. In section 3 we explain the technical details of feature extraction and the video recommendation process as well as Restricted Boltzmann Machines. In section 4 we discuss the developed evaluation methodology and in section 5 we present the obtained results. Finally, in section 6 we provide an overall discussion on the results and in section 7, we conclude the article and list plans for future works.

2 | RELATED WORK

2.1 | Content-Based Recommender Systems

Early approaches in recommender systems were based on the popular Content-based Filtering (CBF) algorithms. These algorithms model user profiles by associating their preferences with the item content Degemmis, Lops, and Semeraro (2007); Eirinaki, Vazirgiannis, and Varlamis (2003); Hawashin, Lafi, Kanan, and Mansour (2019); Jannach et al. (2010); Magnini and Strapparava (2001); Martins, Belém, Almeida, and Gonçalves (2016); Renckes, Polat, and Oysal (2012). The user preferences can be of different forms, e.g., ratings or interactions, and can be elicited *explicitly* Billsus and Pazzani (1999), or *implicitly* Kelly and Teevan (2003). The item content can be represented with different forms of semantic attributes, e.g., item category or item description. These attributes were used by the recommender systems to establish a *Vector Space Model* Pazzani and Billsus (2007a). Accordingly, every item is represented as a multi-dimensional vector associated with the content attributes. This allowed the systems to compute the relevance of user preferences with respect to the item attributes Lops et al. (2011); Vinagre, Jorge, and Gama (2018).

Various CBF algorithms have been used by recommender systems. One of the classical algorithms is *K-Nearest Neighbors (KNN)*. This algorithm computes the similarities of items based on their content attributes, and then recommends, to a target user, the items that are similar to those she liked in the past. The similarity is typically measured based on *Cosine* similarity Lops et al. (2011); Pazzani and Billsus (2007a).

Other content-based algorithms take advantageous of techniques based on *Relevance Feedback* Ahn, Brusilovsky, Grady, He, and Syn (2007) or *Probabilistic Models* Mooney and Roy (2000). For example, by analyzing the item attributes and user preferences, the algorithm could calculate the probability that a target user is interested to a particular item, and then recommend items with highest probabilities.

2.2 | Content-Based Video Recommender Systems

In video recommender system, CBF has been one of the popular approaches. Different types of content attributes used by recommender systems could vary from traditional *semantic* attributes to novel *visual* features. The former type is more *high-level* and it is obtained from traditional sources, ranging from databases, or ontologies, to review websites, or social media Ahn et al. (2007); Billsus and Pazzani (2000); Cantador, Szomszor, Alani, Fernández, and Castells (2008); Middleton, Shadbolt, and De Roure (2004); Mooney and Roy (2000); Musto et al. (2012). The latter type, on the

other hand, is more *low-level* and it is obtained by directly analyzing the video files Deldjoo, Elahi, Quadrana, and Cremonesi (2015); Deldjoo, Elahi, Quadrana, Cremonesi, and Garzotto (2015).

While majority of related works in content-based video recommender systems have focused on using the traditional high-level attributes, a very limited works have investigated the potential of low-level visual features. These features are more representative of the production style and hence they may enable recommender systems to be more *style-aware* Canini, Benini, and Leonardi (2013); Lehinevych et al. (2014); Yang et al. (2007); Zhao et al. (2011). Yang et al. (2007) proposed a *VideoReach* recommender system that exploits a set of of semantic attributes combined with visual features. They have shown that this combination improved the click-through-rate. Zhao et al. (2011) proposed a learning algorithm that integrates multiple ranking lists, each generated by using different feature types, including visual features, and reported improvement over their baseline.

It should be noted that, our work differs from the mentioned related works in various aspects. First of all, these works considered a simplified evaluation methodology with a single scenario. We, on the other hand, have designed and developed a comprehensive evaluation methodology with various realistic scenarios, i.e., *extreme* item cold start, *moderate* item cold start, and item *warm* start Elahi, Ricci, and Rubens (2013); Rubens et al. (2015).

Moreover, these works have only considered using the combination of traditional attributes with visual features. We have considered using the visual features, individually or in combination with traditional attributes. Furthermore, these works have only used a single recommender algorithm while we implemented two different recommender algorithms, i.e., Pure CBF and Hybrid FM (Factorization Machines). Finally, we have developed an artificial neural network based on Restricted Boltzmann Machines (RBM) for its various benefits, one of which being the automatic feature engineering. This type of feature engineering solutions has not been considered in any of the above-mentioned related works.

2.3 | Restricted Boltzmann Machines in Recommender Systems

Algorithms based on Restricted Boltzmann Machine (RBM) as a powerful Artificial Neural Networks (ANN) Khamparia and Singh (2019) have been proved to be effective tools for being integrated in recommender systems. In Salakhutdinov, Mnih, and Hinton (2007), an RBM model was used for recommendation. Van den Oord, Dieleman, and Schrauwen (2013) proposed a method based on ANN in order to learn feature embedding for multimedia recommendation. This embedding has been then used to regularize Matrix Factorization in the recommendation process. Georgiev and Nakov (2013) utilizes user-user and item-item correlations in a recommender algorithm adopting an RBM core. Finally, Sainath, Kingsbury, Sindhwani, Arisoy, and Ramabhadran (2013) implemented matrix factorization and a deep network of ANN in order to decrease the number of the parameters of the model.

Our work is not comparable with these related works as they are not content-based algorithms. Indeed, they used RBM in the recommender systems that are based on Collaborative Filtering (CF) mechanism. And hence, it has been mainly used to for dimensionality reduction in collaborative filtering where dataset contains millions of items or millions of users. In this sense, it could be compared with Matrix Factorization algorithms Koren, Bell, and Volinsky (2009).

However, our approach is a fundamentally different from these collaborative filtering-based approaches as they solely relies on user ratings and not on the item content, when generating recommendations.

3 | TECHNICAL DETAILS

In this research work, we have designed a system with different components, i.e., Video Analyzer, Neural Network based on RBM, and Recommender System (see figure1). The operations performed by these components are listed in the following:

- **Video Analysis** (figure 2)
  - *Movie Segmentation*: every movie is segmented into *shots*, i.e., sequences of consecutive frames captured without interruption of the camera;
  - *Key-Frame Detection*: within every shot the middle frame is selected as representative of the shot (key-frame);
  - *Feature Extraction*: every key-frame is analyzed and the visual features are extracted;
  - *Feature Aggregation*: the feature vectors are aggregated over the entire movie to form a feature vector, descriptive of the whole movie;
- **Training Neural Network (RBM)**
  - a neural network based on Restricted Boltzmann Machines (RBM) is trained on the aggregated visual feature vectors.
- **Recommendation**

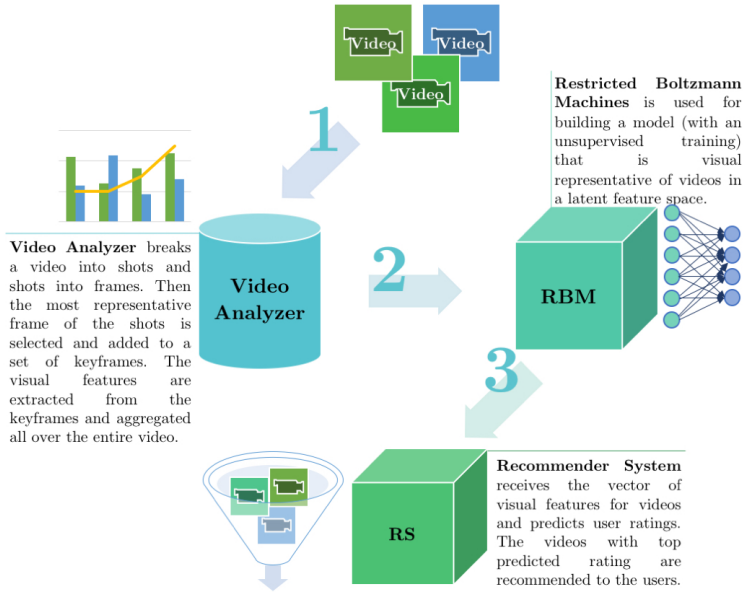

FIGURE 1 The general schema of the developed system, consisted of (i) Video Analyzer component, (ii) Neural Network based on Restricted Boltzmann Machines (RBM), and (iii) Recommender System

- *Pure Content-based Filtering (CBF)* the output layer of the neural network based on RBM (as latent representation of the item content) is used to compute *item-item* similarities and recommendation is generated purely based on these content-wise similarities;
- *Hybrid Factorization Machines (FM)* the hybrid prediction model is trained on (i) the output layer of the neural network based on RBM (as latent representation of the item content), and, (ii) user preference vectors (i.e., user ratings) and recommendation is generated based on the prediction of the trained hybrid model.

3.1 | Feature extraction

We have considered 2 different set of novel *visual* features, i.e., **Mise-en-scène** and **MPEG7** features, that can be extracted automatically by analyzing a video content. Both of these features have been shown excellent power in representing video content Deldjoo, Elahi, Cremonesi, Garzotto, and Piazzolla (2016); Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016); Deldjoo, Elahi, Quadrana, Cremonesi, and Garzotto (2015). In addition, we considered 2 traditional *semantic* attributes, i.e., **Tag** and **Genre**, that requires either crowd-annotation or expert-labeling.

We have also merged the visual features with the semantic attributes and formed a **Combined** vector, representing both visual and semantic aspects of the video content. We refer to this extended vector as *combined* form of extraction, as it requires both *automatic* analysis and *manual* annotation. This has allowed us to investigate whether or not the combination of these two different representation of the video content could result in improvement in the quality of the recommendations.

The Mise-en-scène features are described in the following:

- **Mise-en-scène Visual Features**
  - *Object Motion* descriptor of a video feature can be computed based on the optical flow Barron, Fleet, and Beauchemin (1994); Horn and Schunck (1981) as a robust estimation of velocities within video key-frames. For a key-frame t, we denote $\overline{m}_t$ as the average motion of pixels and $(\sigma^2_m)_t$ as the standard deviation of pixel motions. After computing these values for every key-frame, they are aggregated all over the entire key-frames of a movie Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016):

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

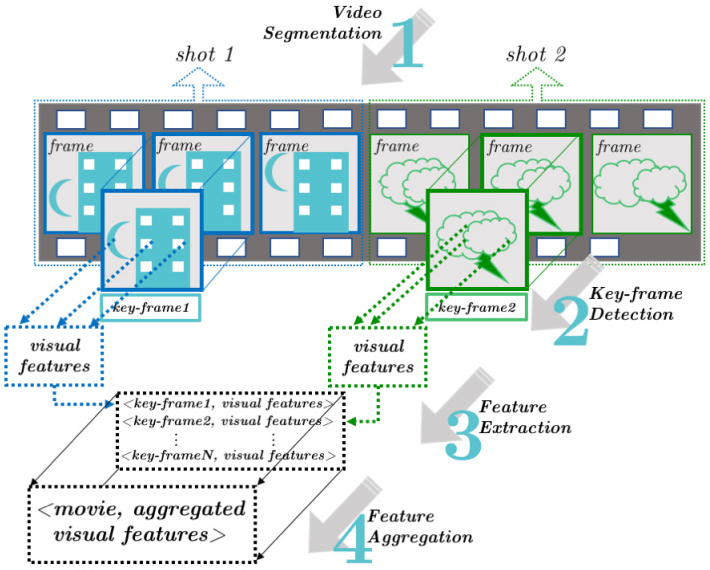

FIGURE 2 Different operations performed by the Video Analyzer component, i.e., (1) Video Segmentation, (2) Key-frame Detection, (3) Feature Extraction, and (4) Feature Aggregation

and

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

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.

– *Color Variance* is measured for each key-frame in the *Luv* color-space by computing the covariance matrix ρ Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016):

$$\rho = \begin{pmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{pmatrix} \tag{3}$$

where $\sigma_L, \sigma_u, \sigma_v, \sigma_{Lu}, \sigma_{Lv}, \sigma_{uv}$ are the standard deviation over three channels L, u, v and their mutual covariance and hence $\Sigma = \det(\rho)$ is a measure of color variance. The mean and standard deviation of Σ over keyframes are aggregated for the entire video:

$$\mu_{cv} = \frac{\sum_{q=1}^{n_{sh}} \Sigma_q}{n_{sh}} \tag{4}$$

where n_{sh} is the number of shots and again it is equal to the number of key-frames.

– *Lighting Key* is measured by transforming the video pixels to *HSV* color-space where the mean μ and standard deviation σ, as representative of to the brightness, are computed. Then the lighting key ξ is measured: Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016):

$$\xi = \mu.\sigma \tag{5}$$

The computed lighting-key of each key-frame is then aggregated all over the entire video.

– *Shot Length* is computed with the following formula Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016):

$$\overline{L}_{sh} = \frac{n_f}{n_{sh}} \tag{6}$$

where n_f is again the total number of frames and n_{sh} the number of shots in a video. Due to differences in frame rates of the videos, $\overline{L}_{sh}$ should be normalized by the frame rate of the video.

We have also considered another type of visual features, based on MPEG7 standard. The MPEG7 features specify several visual descriptors that can be classified into categories of (e.g.,) color-based and texture-based features Manjunath et al. (2001). These features have shown to be powerful in the expressing the color and texture for computing video similarities for both Recommender Systems Deldjoo, Elahi, Quadrana, and Cremonesi (2018) and Information Retrieval applications Manjunath et al. (2001); X.-Y. Wang, Zhang, and Yang (2014). The following list, describes the considered MPEG7 features:

- **MPEG-7 Visual Feature**
  - *Scalable Color Descriptor (SCD)* is measured by computing the color histogram of a video frame in the *HSV* color space. More precisely, SCD feature set can be interpreted as Haar transformation of a color histogram in the *HSV* color space Manjunath et al. (2001). The color histogram bins are also normalized and mapped into a 4-bit integers. This results in giving higher importance to small values. The Haar transform is applied to the 4-bit integer values across the histogram bins. The sums of every two adjacent histogram bins out of 256-bins, can be computed resulting in a representation of a 128-bin histogram. This process can be further repeated, resulting in histograms with 64, 32 or 16 bins Ohm, Kim, and Krishnamachari (2005). However, in our experiments, we still chose to set the number of histogram bins to 256.
  - *Color Structure Descriptor (CSD)* is measured by counting the number of times a color is contained within part of a key-frame Manjunath et al. (2001). This is indeed a modified version of the SCD histogram that considers the physical position of every color within the frames, and hence it captures color space and information on the structure of this space. The CSD feature set contains 256 values and there values are typically not similar and can be used well in the recommendation or retrieval applications. Suppose $c_1, c_2, c_3, ..., c_M$ denote the M colors within a key-frame. The CSD feature set is then computed as Manjunath et al. (2001); Ohm et al. (2005):

$$csd = h(m), \;\; m \in \{1, ..., M\} \tag{7}$$

    where $h(m)$ is the color histogram. A bin of the color histogram counts the number of times a pixel with a color m is found within the key-frame. Here m is basically the index of colors in which the key-frame is represented. M is the number of histogram bins, and can be chosen from the set 32, 64, 128, 256. In our experiment, the number of histogram bins is set to 256.
  - *Color Layout Descriptor (CLD)* is measured by applying the *Discrete Cosine Transform (DCT)* on color vector in YC_BC_R color space, where Y represents *luminance* factor, C_B the *blue* factor and and C_R represents the *red* factor Haskell, Puri, and Netravali (1996); Manjunath et al. (2001). CLD is indeed a powerful set of features that is resolution invariant and it can reflect the spatial distribution of colors within a key-frame and can be adopted for various applications with recommendation and retrieval goals Ohm et al. (2005). In our experiment, the length of this set is 120.
  - *Edge Histogram Descriptor (EHD)* is measured by computing the local *edge distribution* in a key-frame. The key-frame is divided into 16 non-overlapping blocks (sub-frames). Edges within each block are classified into the following categories: vertical, horizontal, left diagonal, right diagonal and non-directional edges. The final local edge descriptor encompasses a histogram with 80 histogram bins.
  - *Homogeneous Texture Descriptor (HTD)* is measured by computing the homogeneous texture regions within a key-frame, by using a vector of 62 energy values. The HTD features represent a quantitative characterization of texture within key-frames. First a set of orientation and scale sensitive filters are applied to a key-frame and then the mean and standard deviation for the output of the filtered key-frame are computed Manjunath et al. (2001). It should be noted that, the procedure, described above, is conducted in frequency domain rather than spatial domain since experiments have shown that, in this way, the computational complexity of extracting HTD features drops substantially Manjunath et al. (2001). Prior works have reported the robustness and effectiveness of HTD features in representing video content Haley and Manjunath (1999); Ma and Manjunath (1998).

3.2 | Training the Neural Network (RBM)

We have designed and developed a neural network model based on Restricted Boltzmann machines (RBM). The network has been trained on the visual features (as input layer) and then the output layer of the network is used as representative of the video items. This leads to an improved representation of the video content. Indeed, prior works have shown that visual features are not equally representative and informative of the video content Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016). RBM is capable of weighting the visual features according to their informativeness, in a completely automatic way. RBM is known to be a *generative* model as it can map any input data to a different output space, hence generating it. This process may result in an enhanced representation of data, and hence, RBM can be used for dimensionality reduction, noise removal, and feature engineering.

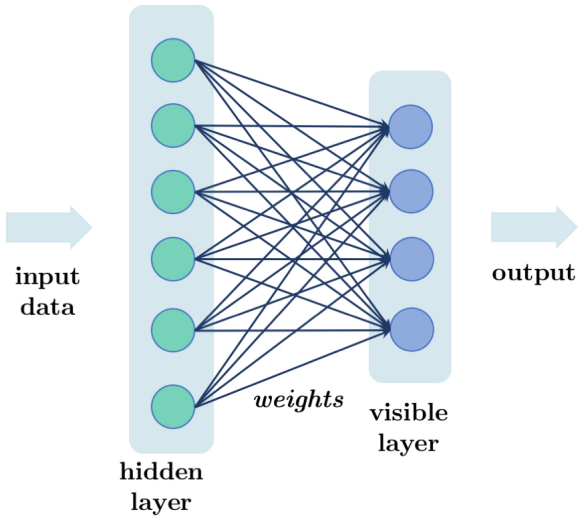

FIGURE 3 The structure of the neural network based on Restricted Boltzmann Machines (RBM). Any type of features can be given as input data to the RBM network. The network then learns how useful is each of the features and weights them according to that factor. The output layer is used by the content-based recommender algorithm to generate personalized recommendations for users.

The structure of our RBM network is illustrated in Figure 3. As it can be seen in that figure, RBM consists of a *visible layer (v)* and a *hidden layer (h)* Hinton and Salakhutdinov (2006). RBM operates by computing the probability distribution over a set of hidden variables within the hidden layer, through solving a set of differential equations.

Formally, RBM can be represented by a bipartite network, comprised of units, each of which containing (stochastic) binary neurons. Each neuron can be in either of 2 states, i.e., in *active* or *inactive* state (hence, $v, h \in \{0,1\}$). The activation of neurons is done through the smooth *Sigmoid function*. Because of independence of nodes in the same layer, the conditional probability of the hidden layer units are independent from each other and they are calculated as:

$$p(h|v) = \prod_{i=1}^{p} p(h_i|v) \tag{8}$$

where the conditional probability that one hidden node is set to 1 is calculated as

$$p(h_i = 1|v) = \sigma(\sum_{j=1}^{q} w_{ij} v_j + b_i) \tag{9}$$

where the w_{ij} are the learned *weights* and b_i term is the bias. RBM learns the values of hidden layer (h_i), that in turn generate the values of visible layer (v_i), by optimizing the log-likelihood of training data as Eq.10.

$$l(\theta) = \frac{1}{n} \log p(v) = \frac{1}{n} \sum_{i} \log(p(v_i)) \tag{10}$$

where $p(v)$ can be calculated through its marginal distribution over its joint probability to values of hidden layer as in Eq.11

$$p(v) = \sum_{h} p(v, h) \tag{11}$$

The joint probability of v and h are defined as:

$$p(v, h) = \frac{exp(-\mathcal{E}(v, h))}{z} \tag{12}$$

where z is the normalization constant and $\mathcal{E}$ is the energy function defined as:

$$\mathcal{E} = -h^T w v - b^T h - c^T v \tag{13}$$

The activation of both visible layer (v) and hidden layer (h) is calculated by sampling from $p(h|v)$ and $p(v|h)$.

3.3 | Recommendation algorithm

We have used two different recommender algorithms, both capable of incorporating item content into recommendation generation process. The first algorithm is *Pure Content-based Filtering (Pure CBF)*, that computes recommendations solely based on the item content, while the second algorithm, *Hybrid Factorization Machines (Hybrid FM)*, extends Matrix Factorization (MF), and in addition to the user ratings, takes into account the item content when computing recommendations.

3.3.1 | Pure CBF

is a traditional algorithm for generating recommendation. This algorithm computes item-item similarities of videos and uses them in the recommendation process. The similarity is computed based on *Cosine* similarity. Suppose that $\mathcal{U}$ is the set of all users and $\mathcal{I}$ is the set of all items, and $\mathcal{R}$ is the rating matrix, with rows being users and columns being items and matrix entries corresponding to the rating preferences. Every item $j \in \mathcal{I}$ can be presented by a feature vector f_j associated with that item. Similarity score for a pair of items j and j′ with feature vectors f_j and $f_{j'}$, is calculated as Eq. 14.

$$s_{jj'} = \frac{f_j^T f_{j'}}{||f_j||||f_{j'}||} \tag{14}$$

Based on the computed similarities between an item j with all other items, a set of *K-Nearest Neighbors (*NN_j*)* is found comprising the most similar items to the item j. The ratings that have given to these most similar items are used to compute rating prediction for item j.

Suppose $\hat{r}_{ij}$ is the predicted rating that user i may give to the item j. Eq. 15 presents how $\hat{r}_{ij}$ is computed:

$$\hat{r}_{ij} = \frac{\sum_{j' \in NN_j, r_{ij'}>0} r_{ij'} s_{sjj'}}{\sum_{j' \in NN_j, r_{ij'}>0} s_{sjj'}} \tag{15}$$

where $r_{uj} > 0$ denotes the not-null entries within the rating matrix $\mathcal{R}$, i.e., user ratings already available in the rating dataset. In our experiments, the number of nearest neighbors (K) is set to 100.

3.3.2 | Hybrid FM

is the state-of-the-art recommender algorithm that extends the classical Matrix Factorization (MF) algorithm Koren et al. (2009). Classical MF can learn latent factors for each user and item based on the available (known) ratings. Then this algorithm exploits learned latent factors in order to predict unknown ratings. Hybrid FM extends the classical FM and combines it with the *Support Vector Machines (SVM)*, which enables it to adopt not only the user ratings, but also any other form of data associated with users and items (as known as *side information*) Low et al. (2012); Rendle (2012a); TURI (2018). This includes semantic attributes (e.g., genre and tag) as well as visual features (e.g., Mise-en-scène and MPEG7). When the model behind hybrid FM is trained on known user ratings and item features, it can predict the unknown ratings Rendle (2012a). For instance, suppose a user likes comedy videos and dislikes action videos. The model can capture that pattern and consequently predict lower ratings for action videos and higher ratings for comedy videos.

Formally, the predicted rating of a user i for an item j is computed by

$$\hat{r}_{ij} = \mu + w_i + w_j + \mathbf{a^T x_i} + \mathbf{b^T y_i} + \mathbf{u_i}^\mathbf{T} \mathbf{v_j} \tag{16}$$

where μ represents the global bias term, w_i is the weight term for a user i, w_j is the weight term for an item j, $\mathbf{x_i}$ and $\mathbf{y_j}$ are the user and item feature vectors, respectively. The terms $\mathbf{a}$ and $\mathbf{b}$ are the weight vectors for user and item features. The latent factors for the user and item are given by $\mathbf{u_i}$ and $\mathbf{v_j}$ (similar to latent factors in classical MF). We should note that, this algorithm can also learn user features (e.g., user age and gender). However, this is not the focus of this work and hence we have not considered this type of user features.

Finally, the model is trained by optimizing the following objective function Rendle (2012a); TURI (2018):

$$\min_{\mathbf{w},\mathbf{a},\mathbf{b},\mathbf{V},\mathbf{U}} \frac{1}{|\mathcal{R}|} = \sum_{(i,j,r_{ij}) \in \mathcal{R}} \mathcal{L}(\hat{r}_{ij}, r_{ij}) + \lambda_1(||\mathbf{w}||_\mathbf{2}^\mathbf{2} + ||\mathbf{a}||_\mathbf{2}^\mathbf{2} + ||\mathbf{b}||_\mathbf{2}^\mathbf{2}) + \lambda_2(||\mathbf{U}||_\mathbf{2}^\mathbf{2} + ||\mathbf{V}||_\mathbf{2}^\mathbf{2})$$

where $\mathcal{R}$ is the rating dataset, r_{ij} is the rating that user i gave to item j, $\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, ...)$ denotes the user's latent factors and $\mathbf{V} = (\mathbf{v_1}, \mathbf{v_2}, ...)$ denotes the item latent factors. λ_1 denotes the linear regularization parameter and λ_2 the regularization parameter. *Stochastic Gradient Descent (SGD)* is adopted for the training.

In our experiments, we have adopted the Hybrid FM implemented in *GraphLab* package in our experiments Low et al. (2012); TURI (2018).

4 | EVALUATION METHODOLOGY

4.1 | Dataset

Our dataset contains 3'964 movie trailers, collected by querying *YouTube* dataset Harper and Konstan (2016). It has been shown that the movie trailers are visually very similar and representative of to their corresponding full-length movies Deldjoo, Elahi, Cremonesi, Garzotto, Piazzolla, and Quadrana (2016).

The number of ratings associated with these movie trailers is 8'931'665 ratings, given by 242'209 users to these movie trailers. The sparsity of the rating dataset is about 99.06%.

The Movielens dataset also contains the 586'994 tag attributes to the queried movie trailers and classifies them into 19 genre attributes, i.e., *action*, *adventure*, *animation*, *children's*, *comedy*, *crime*, *documentary*, *drama*, *fantasy*, *film-noir*, *horror*, *musical*, *mystery*, *romance*, *sci-fi*, *thriller*, *war*, *western*, and *unknown*. Each movie trailer is labeled with a single or multiple genre attribute(s).

4.2 | Evaluation Scenarios

We have designed three realistic evaluation scenarios, i.e., *Extreme* item cold start, *Moderate* item cold start, item *Warm* start scenario (see figure 4).

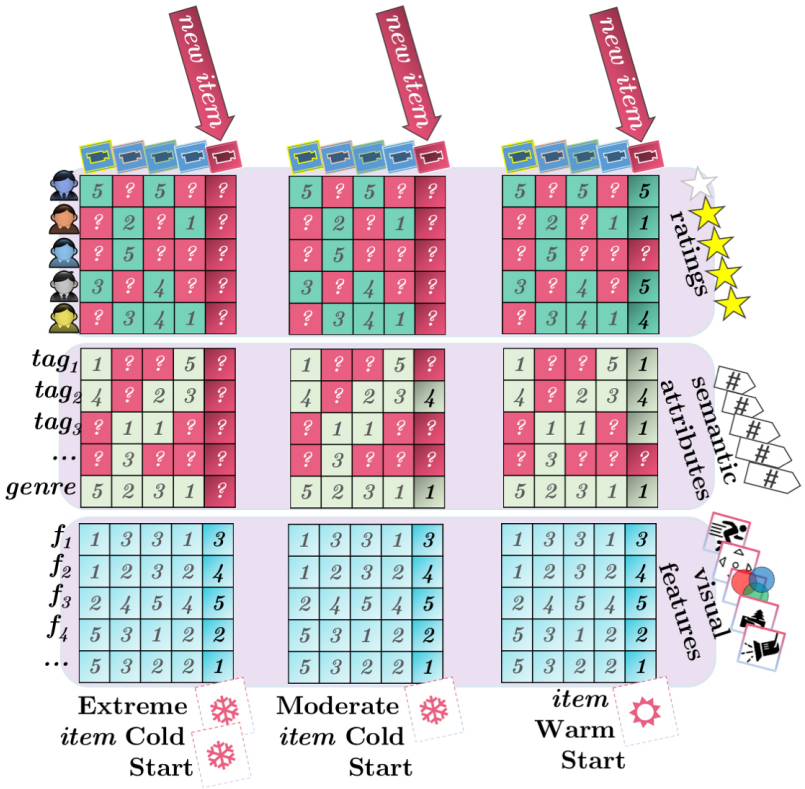

FIGURE 4 Different realistic scenarios that can occur to any real-world recommender system, depending on the level available user-annotated data. These scenarios includes (i) Extreme item Cold Start, (ii) Moderate item Cold Start, and (iii) item Warm Start. Within each scenario, certain values of user-annotated attributes can be unknown to the recommender system. These missing attributes are indicated with "?" mark.

4.2.1 | Extreme Item Cold Start

in recommender systems occurs when a video is added to the item catalog without any data that can describe that item. Consequently, the system would fail to recommend that item to any user. This is a serious problem mainly in video-sharing applications (e.g., YouTube) where every hour

more than 500 hours of video content is added to the system catalog Insights (2018). This can be also considered as *New System* problem which refers to a new recommender system in its early stage and it has not been used by many users Karimi, Nanopoulos, and Schmidt-Thieme (2015). Hence, the dataset is very limited and almost no item has received explicit user annotation, e.g., in terms of tagging or ratings. To simulate this severe scenario, we assumed that the videos have absolutely no human-annotated data (i.e., tag and genre attributes). Hence, we removed these semantic attributes from the items within the dataset. It is worth noting that, in this extreme scenario, the system can still analyze the video files automatically and extract a set of visual features to represent the video items.

4.2.2 | Moderate Item Cold Start

happens when the semantic attributes are partially available, e.g., for some of the items. We have simulated this scenario by randomly selecting 500 videos and removing the associated content attributes for these videos. Hence, moderate cold start can be considered as a mixed scenario of extreme cold start for some videos and warm start for the rest. Hence, it can be seen as an intermediate situation when a recommender system is in a transition phase from extreme cold start to warm start situation. However, this still means that the system has a serious problem as the items in the extreme cold start situation may not be included to the recommendations for users due to lack of describing data. This also avoids these items to obtain any user preferences which makes the situation even worse. These items may include new items which can be very interesting to users and exclusion of them may reduce novelty of the recommendations and negatively affect the user satisfaction. Since automatically-extracted visual features do not require any human involvement, they will be available the recommender systems within this scenario.

4.2.3 | Item Warm Start

can be considered as the best possible scenario for recommender systems as the items have already obtained considerable user-annotated data that can be well exploited for recommendation. To simulate this scenario, we have included the entire set of semantic attributes (i.e., genre and tag) in the dataset. Similar to the previous cold start scenarios, the visual features can be extracted and exploited by recommender systems in warm start situation. Hence, in addition to the semantic attributes, in this evaluation scenario, we have included the visual features to the dataset.

4.3 | Evaluation Setup

We have implemented an evaluation setup for testing the Top-N recommendation quality.

- We employed *k-fold cross validation* by randomly splitting the dataset into 5 non-overlapping subsets, where in every iteration, $\frac{4}{5}$ of instances are used as training the models and the rest $\frac{1}{5}$ for testing.
- We measure the quality of the recommendation in terms of different evaluation metrics, commonly used in recommender systems research field, i.e., *Normalized Root Mean Square Error (NRMSE)*, *Precision* and *Diversity* Schedl, Zamani, Chen, Deldjoo, and Elahi (2017). In all experiments the recommendation size N (*cut-off*) is set to 5.

NRMSE is calculated by measuring the mean of the squared *rating prediction error*, i.e, the deviation of predicted ratings and true ratings within the test set. Then the squared root of this value is computed and normalized to be in the range of 0 to 1. **Precision** is computed by measuring the percentage of relevant items (i.e., items with rating 4 and 5) within the recommendation list. Since the size of recommendation list is 5 this metric is indeed *precision@5*. **Diversity** is computed by measuring the intra-list dissimilarity between items within a recommendation list, similar to the approach in Elahi et al. (2017). We computed the pairwise cosine similarity of videos, recommended to a user and calculated the mean similarity S. Then diversity is computed as the complement of this intra-list similarity as $(1 - \mathrm{S})$. For the sake of simplicity, all metrics are normalized and presented in the range 0-100%.

4.3.1 | Parameters

In RBM model, the number of hidden units is set to 100 and learning rate is set to 0.1. Number of epochs are 500 and the batch size is set to 50. In the Pure CBF recommendation algorithm, K (number of nearest neighbors) is set to 100. In Hybrid FM, the number of factors are set to 32. For running the experiments, we have used an AMAZON AWS (computation-optimised) server with a pre-installed Ubuntu 14.04 (linux-aws kernel). The server had the following hardware specifications: vCPU: 16, ECU: 62, RAM: 30 GiB, EBS drive: 100 GiB.

5 | RESULTS

In this section, we describe the results obtained from several experiments, considering 3 evaluation scenarios, i.e., *extreme* (item) cold start, *moderate* (item) cold start, and (item) warm start (see figure 4).

5.1 | Extreme Item Cold Start

Table 1 presents the results for the *extreme* item cold start scenario. As it can be seen, in this scenario, when the recommendations are generated by the content-based algorithm (pure CBF), MPEG7 features achieve the best results in terms of NRMSE and precision values, i.e., 22.3 and 40.6, respectively. In terms of diversity, Mise-en-scène features obtain the highest score of 71.9. When hybrid FM (Factorization Machines) algorithm is used, the recommendation based on visual features substantially outperforms the recommendation based on semantic attributes (i.e., tag and genre baselines). In terms of NRMSE, Mise-en-scène features obtains the lowest error value of 18.1, while in terms of precision both Mise-en-scène and MPEG7 features perform excellent and obtain the value of 69.9. In terms of diversity, again both Mise-en-scène and MPEG7 features obtained the best value of 70.7.

These results indicate the clear superiority of the recommendation based visual features in the extreme cold start scenario. Indeed, in some cases the improvement of visual features over tag features (the stronger baseline) reaches 55.0% for NRMSE (hybrid FM trained on Mise-en-scène features) and 82.9% (hybrid FM trained on either Mise-en-scène or MPEG7 features). This is interesting as it shows the power of visual features that can be extracted *automatically* in comparison to traditional attributes that need *manual* human annotation.

It is worth noting, that even small improvement of metrics such as NRMSE is pretty tough. For instance, the winner team of *Netflix 1 Million USD Prize* , achieved %10.06 improvement on rating prediction task, after years.

Since in this scenario, non of the traditional attributes are available, the combination of these attributes with visual features is Not Applicable and presented with “NA" in the table 1.

5.2 | Moderate Item Cold Start

Table 2 shows the results of *moderate* item cold start scenario. In this scenario, limited amount of semantic attributes (i.e., tags and genre) is available and can be used by the recommender system. Consequently, although the quality of recommendations based on these attributes have been considerably improved, however, the best performance has been achieved by using the *combined* features, i.e., combination of visual features with traditional attributes. Accordingly, two similar combinations of Mise-en-scène + tag + Genre and Mise-en-scène + tag outperforms other baselines by reaching the NRMSE and precision of 18.1 and 46.5. In terms of diversity, the highest value belongs to Mise-en-scène features with the value of 71.9.

In the case of Hybrid FM recommender algorithm, still individual visual features present the best results. The recommendation based on Mise-en-scène features obtains the lowest NRMSE value of 18.1. In terms of precision, both Mise-en-scène and MPEG7 features overtake all other feature sets by getting the highest value of 69.9. For diversity, the tag attributes obtains the best value, i.e. 72.1.

The results of moderate item cold start scenario is also very interesting as they show that although the quality of recommendation based on traditional semantic attributes have been boosted, still recommendations based on visual features can be of higher quality.

It has to be noted that although due to availability of certain amount of semantic attributes, the recommendations based on these attributes have been improved, however, still these attributes are collected with a cost of crowd annotation (for tag attributes) and expert labeling (for genre attributes). Visual features, on the other hand, does not require any of these manual human annotation and can be extracted completely automatic. Indeed, despite the fact that visual features outperform the traditional features, even observing a similar performances could still be considered as interesting results.

5.3 | Item Warm Start

In Table 3 the results of the experiments in the item warm start scenario have been presented. This is a scenario where large amount of semantic data (i.e., tags and genres) is available for the items. As expected, in this scenario, the results of recommendations based on semantic attributes shows improvement over the results of previous moderate and extreme cold start scenarios. However, again, the best performance has been achieved by recommendation with visual features, either used individually or combined with semantic attributes.

In case of Pure CBF recommender algorithm, combination of visual features with traditional attribute, i.e., Mise-en-scène + tag and Mise-en-scène + tag + Genre, have obtained the best results. The recommendation based on the former combination achieved NRMSE and precision of 18.0

TABLE 1 Comparison of recommendations based on visual features (extracted automatically) and semantic attributes (annotated manually) in the **Extreme item Cold Start** scenario. The comparison is made with two different content-based recommender algorithms, i.e., Pure Content-Based Filtering and Hybrid Factorization Machines. Due to the particular setting of this scenario, combining visual features and semantic attributes is not applicable in this scenario and hence it is presented with "NA"

Recommender	Extraction	Feature/Attribute	NRMSE %	Precision %	Diversity %
Pure CBF	*manual*	Tag	40.0	38.2	70.3
	manual	Genre	40.3	37.6	69.5
	automatic	MPEG7	**22.3**	**40.6**	71.8
	automatic	Mise-en-scène	22.4	40.0	**71.9**
	combined	Mise-en-scène + Tag	NA	NA	NA
		Mise-en-scène + Genre	NA	NA	NA
		Mise-en-scène + Tag + Genre	NA	NA	NA
		MPEG7 + Tag	NA	NA	NA
		MPEG7 + Genre	NA	NA	NA
		MPEG7 + Tag + Genre	NA	NA	NA
Hybrid FM	*manual*	Tag	40.0	38.2	70.3
	manual	Genre	40.3	37.6	69.5
	automatic	MPEG7	18.7	**69.9**	**70.7**
	automatic	Mise-en-scène	**18.1**	**69.9**	**70.7**
	combined	Mise-en-scène + Tag	NA	NA	NA
		Mise-en-scène + Genre	NA	NA	NA
		Mise-en-scène + Tag + Genre	NA	NA	NA
		MPEG7 + Tag	NA	NA	NA
		MPEG7 + Genre	NA	NA	NA
		MPEG7 + Tag + Genre	NA	NA	NA

and 46.6, respectively. The latter combination shows excellency only in terms of NRMSE with the similar value of 18.0. With respect to diversity, the highest value is for to Mise-en-scène features, i.e., 71.9.

When the Hybrid FM recommender algorithm is adopted, the visual features, used individually, outperform the baselines. The highest recommendation quality is obtained by Mise-en-scène features with NRMSE value of 18.1 and precision of 69.9. Similarly MPEG7 features obtain highest precision value 69.9. With respect to diversity, the combination of tag attributes with MPEG7 features shows the highest value of 71.8.

As noted before, warm start scenario illustrates a situation where a recommender system have elicited substantial amount of data associated with tradition semantic attributes. Hence, the quality of recommendations based on these attributes may reach its maximum level. In such a scenario, competing with these manually-annotated semantic attributes may not be an easy task for automatically-extracted visual features. Indeed, this gets interesting when knowing that for one type of considered attributes (i.e., genre) a group of video experts has been involved to carefully label the video items and for the other type (i.e., tags) a large community of users have been involved in annotating the video items. However, still our results confirm that proper engineering of the visual features with the neural network and adopting the state-of-the-art recommender algorithms based on sophisticated Machine Learning methods can lead to the superior performance.

6 | DISCUSSION

Our comprehensive experiments conducted in three different scenarios have addressed the research questions we have formulated:

- **[RQ1]** Can recommendation based on the visual features, extracted automatically, remedy the *New Item* problem in *extreme cold start* scenario;
- **[RQ2]** Can recommendation based on the automatic visual features effectively remedy the *New Item* problem in *moderate cold start* scenario;

TABLE 2 Comparison of recommendations based on visual features (extracted automatically) and semantic attributes (annotated manually) in the **Moderate item Cold Start** scenario. The comparison is made with two different content-based recommender algorithms, i.e., Pure Content-Based Filtering and Hybrid Factorization Machines.

Recommender	Extraction	Feature/Attribute	NRMSE %	Precision %	Diversity %
Pure CBF	*manual*	Tag	18.5	46.2	70.6
	manual	Genre	20.3	42.6	66.0
	automatic	MPEG7	22.3	40.6	71.8
	automatic	Mise-en-scène	22.4	40.0	**71.9**
	combined	Mise-en-scène + Tag	**18.1**	**46.5**	71.2
		Mise-en-scène + Genre	20.2	42.6	67.4
		Mise-en-scène + Tag + Genre	**18.1**	**46.5**	70.8
		MPEG7 + Tag	18.9	42.1	71.5
		MPEG7 + Genre	20.7	42.1	69.2
		MPEG7 + Tag + Genre	19.0	42.4	69.2
Hybrid FM	*manual*	Tag	21.2	67.9	**72.1**
	manual	Genre	23.7	61.2	69.2
	automatic	MPEG7	18.7	**69.9**	70.7
	automatic	Mise-en-scène	**18.1**	**69.9**	70.7
	combined	Mise-en-scène + Tag	20.1	68.8	69.9
		Mise-en-scène + Genre	21.0	65.1	70.1
		Mise-en-scène + Tag + Genre	19.8	66.8	69.5
		MPEG7 + Tag	21.6	66.4	71.0
		MPEG7 + Genre	24.2	63.0	70.5
		MPEG7 + Tag + Genre	20.0	65.9	70.0

- **[RQ3]** How combining visual features with traditional semantic attributes can improve the quality of recommendation in *warm start* scenario.

The promising results obtained from these experiments may provide a solid proof that the recommending videos based on visual features (i.e., MPEG7 and Mise-en-scène) not only effectively solve the *extreme* and *moderate* item cold start problem but also improves the quality of recommendation in item *warm* start situation.

Table 4 provides an overall summary of all results obtained from the entire experiments. The star mark(s) in the table indicate(s) the level of performance excellency, within each evaluation scenario and with respect to a certain metric. Accordingly, the best performances are marked with double stars (★★), and the rest, with a single star (★). The worst performances are marked with a cross symbol (×). At the bottom of the table, the overall best feature/attribute is presented, by comparing the results obtained from both of the recommender algorithms, for each evaluation scenario and evaluation metric.

As it can be seen, in all evaluation scenarios and with respect to almost all evaluation metrics, recommendation based on either Mise-en-scène (MISE) or MPEG7 visual features has achieved the best results. The only exception is observed for the diversity metric in moderate item cold start scenario where the recommendation based on tag attributes obtained the best results. In addition to that, in item warm start scenario, in terms of NRMSE, the combination of visual features with semantic attributes (Mise-en-scène + Tag + Genre) obtained the best results. Comparing the recommender algorithms, overall, we have observed better performance with Hybrid FM than the Pure CBF, especially in extreme cold start situation.

Regardless of the recommender algorithm, the excellent performance of the visual features is very promising particularly by taking into account that these features are extracted from videos automatically, without any need for human-annotation. The semantic attributes, on the other hand, completely depend on human effort whether in the form of crowd-annotation (for tags) or expert labeling (for genre). Such a an extensive human effort can not be always assumed to be easily available.

Moreover, such semantic attributes are typically extremely noisy and sparse (e.g., 96% sparsity for tags). This is while the sparsity of visual features is 0% meaning that the feature matrix is a full matrix. In addition to that, the features are extracted from each key-frame (frame-level

TABLE 3 Comparison of recommendations based on visual features (extracted automatically) and semantic attributes (annotated manually) in the **item Warm Start** scenario. The comparison is made with two different content-based recommender algorithms, i.e., Pure Content-Based Filtering and Hybrid Factorization Machines.

Recommender	Extraction	Feature/Attribute	NRMSE %	Precision %	Diversity %
Pure CBF	*manual*	Tag	18.4	46.5	70.6
	manual	Genre	20.2	42.8	66.1
	automatic	MPEG7	22.3	40.6	71.8
	automatic	Mise-en-scène	22.4	40.0	**71.9**
	combined	Mise-en-scène + Tag	**18.0**	**46.6**	71.1
		Mise-en-scène + Genre	20.2	42.5	67.4
		Mise-en-scène + Tag + Genre	**18.0**	46.5	70.7
		MPEG7 + Tag	19.2	41.7	71.3
		MPEG7 + Genre	20.7	42.2	69.2
		MPEG7 + Tag + Genre	19.2	42.1	70.5
Hybrid FM	*manual*	Tag	19.8	68.6	71.0
	manual	Genre	22.5	61.2	67.5
	automatic	MPEG7	18.7	**69.9**	70.7
	automatic	Mise-en-scène	**18.1**	**69.9**	70.7
	combined	Mise-en-scène + Tag	19.4	69.5	68.8
		Mise-en-scène + Genre	21.5	65.9	69.5
		Mise-en-scène + Tag + Genre	19.4	67.6	70.4
		MPEG7 + Tag	19.7	68.7	**71.8**
		MPEG7 + Genre	23.9	63.2	69.7
		MPEG7 + Tag + Genre	20.9	66.3	69.5

extraction) while the traditional attributes are given to an item (item-level annotation). Hence, the quality of the recommendation can be further improved by designing a more advance aggregation method. As noted before, we have used basic aggregation method, i.e., taking the average of the visual features over the entire video. A more advanced aggregation is among our plans for future work.

Finally, when combining the visual features with semantic attributes, we have simply merged the vectors in order to form a single vector for every video. This could be the reason that the combination of these visual features and semantic attributes could not outperform the performance of the individual visual features. We plan to develop a more advance fusion technique that can result in improvement in the recommendation performance. However, still in some cases, e.g., when there is certain amount of semantic attributes (moderate cold start and warm start), in terms of recommendation accuracy (precision), combining visual features and semantic attributes has achieved the best performance.

7 | CONCLUSION

In this article, we have addressed the *New Item* cold start problem in recommender systems. This problem occurs when a new item is added to the item catalog and no human-annotated data is available for that item. In such a situation any recommender algorithm may fail to recommend this item to the users.

In addressing this problem, this article proposes using a set of novel features that do not require any human annotation, and can be extracted by visually analyzing the video files. In order to evaluate the effectiveness of the proposed visual features, we have designed and developed 3 different realistic scenarios, i.e., (i) *extreme* item cold start, *moderate* item cold start, and item warm start. Each of these scenarios represents a situation that any recommender system may encounter, depending on the availability of the data. We have adopted two different recommender algorithms, both capable of using content features in the recommendation process, i.e. Pure CBF and Hybrid FM.

TABLE 4 Summary of all results obtained from the entire set of experiments, within 3 different evaluation scenarios. The star mark(s) indicate(s) the level of performance excellency, within each evaluation scenario and with respect to a certain metric. The best performances are marked with double stars (★★), and the rest with a single star (★). The worst performances are marked with a cross symbol (×). Due to width limit of the table, Mise-en-scène features are represented as "MISE".

Rec.	Feature / Attribute	Cold Start						Warm Start		
		Extreme			Moderate					
		NRMSE	Precision	Diversity	NRMSE	Precision	Diversity	NRMSE	Precision	Diversity
Pure CBF ★	Tag	★	★	★	★	★	★	★	★	★
	Genre	×	×	×	★	★	×	★	★	×
	MPEG7	★★	★★	★	★	★	★	★	★	★
	MISE	★	★	★★	×	×	★★	×	×	★★
	MISE + Tag	NA	NA	NA	★★	★★	★	★★	★★	★
	MISE + Genre	NA	NA	NA	★	★	★	★	★	★
	MISE + Tag + Genre	NA	NA	NA	★★	★★	★	★★	★	★
	MPEG7 + Tag	NA	NA	NA	★	★	★	★	★	★
	MPEG7 + Genre	NA	NA	NA	★	★	★	★	★	★
	MPEG7 + Tag + Genre	NA	NA	NA	★	★	★	★	★	★
Hybrid FM ★★	Tag	★	★	★	★	★	★★	★	★	★
	Genre	×	×	×	★	×	×	★	×	×
	MPEG7	★	★★	★★	★	★★	★	★	★★	★
	MISE	★★	★★	★★	★★	★★	★	★★	★★	★
	MISE + Tag	NA	NA	NA	★	★	★	★	★	★
	MISE + Genre	NA	NA	NA	★	★	★	★	★	★
	MISE + Tag + Genre	NA	NA	NA	★	★	★	★	★	★
	MPEG7 + Tag	NA	NA	NA	★	★	★	★	★	★★
	MPEG7 + Genre	NA	NA	NA	×	★	★	×	★	★
	MPEG7 + Tag + Genre	NA	NA	NA	★	★	★	★	★	★
Overall Best		**MISE** (visual)	**MISE** or **MPEG7** (visual)	**MISE** (visual)	**MISE** (visual)	**MISE** or **MPEG7** (visual)	**Tag** (semantic)	**MISE + Tag + Genre** (combined)	**MISE** or **MPEG7** (visual)	**MISE** (visual)

The results of our extensive experiments have shown that using visual features can effectively solve the new item cold start problem. Indeed, in all considered scenarios, and with respect to various metrics (i.e., NRMSE, Precision and Diversity), visual features, either used individually or in combination with other attributes, have achieved the best performance in comparison the the baseline attributes (i.e., genre and tags).

For the future work, we plan to extend the set of visual features by extracting more *high-level* features such as facial expressions within the videos. Recent studies have shown the potential of such facial features and their correlations with user preferences Tkalčič et al. (2017). We also plan to develop more advance techniques for modeling videos from the visual features extracted frame-by-frame. We will also work on better fusion methods for combining visual features with semantic attributes.

Conflict of interest

The authors declare no potential conflict of interests.

References

Adomavicius, G., & Tuzhilin, A. (2005). 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.

Aggarwal, C. C. (2016a). Content-based recommender systems. In *Recommender systems* (pp. 139–166). Springer.

Aggarwal, C. C. (2016b). An introduction to recommender systems. In *Recommender systems* (pp. 1–28). Springer.

Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., & Syn, S. Y. (2007). Open user profiles for adaptive news systems: help or harm? In *Proceedings of the 16th international conference on world wide web* (pp. 11–20).

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65.

Anderson, C. (2006). *The Long Tail*. Random House Business.

Apostolidis, E., & Mezaris, V. (2014). Fast shot segmentation combining global and local visual descriptors. In *Acoustics, speech and signal processing (icassp), 2014 ieee international conference on* (pp. 6583–6587).

Balabanović, M., & Shoham, Y. (1997, March). Fab: Content-based, collaborative recommendation. *Commun. ACM*, *40*(3), 66–72. doi: 10.1145/245108.245124

Bao, X., Fan, S., Varshavsky, A., Li, K., & Roy Choudhury, R. (2013). Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *Proceedings of the 2013 acm international joint conference on pervasive and ubiquitous computing* (pp. 197–206).

Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of optical flow techniques. *International journal of computer vision*, *12*(1), 43–77.

Bastan, M., Cam, H., Gudukbay, U., & Ulusoy, O. (2010). Bilvideo-7: an mpeg-7-compatible video indexing and retrieval system. *IEEE MultiMedia*, *17*(3), 62–73.

Beecks, C., Schoeffmann, K., Lux, M., Uysal, M. S., & Seidl, T. (2015). Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments..

Billsus, D., & Pazzani, M. J. (1999). *A hybrid user model for news story classification*. Springer.

Billsus, D., & Pazzani, M. J. (2000). User modeling for adaptive news access. *User modeling and user-adapted interaction*, *10*(2-3), 147–180.

Bogdanov, D., & Herrera, P. (2011). How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In *Ismir* (pp. 97–102).

Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying low-level and high-level music similarity measures. *Multimedia, IEEE Transactions on*, *13*(4), 687–701.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193.

Bovik, A. C. (2010). *Handbook of image and video processing*. Academic press.

Brezeale, D., & Cook, D. J. (2008). Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *38*(3), 416–430.

Buckland, W. (2008). 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.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331-370. Retrieved from `./papers/burke-umuai-ip-2002.pdf`

Canini, L., Benini, S., & Leonardi, R. (2013). Affective recommendation of movies based on selected connotative features. *Circuits and Systems for Video Technology, IEEE Transactions on*, *23*(4), 636–647.

Cantador, I., Bellogín, A., & Vallet, D. (2010). Content-based recommendation in social tagging systems. In *Proceedings of the fourth acm conference on recommender systems* (pp. 237–240).

Cantador, I., Konstas, I., & Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *Web semantics: science, services and agents on the World Wide Web*, *9*(1), 1–15.

Cantador, I., Szomszor, M., Alani, H., Fernández, M., & Castells, P. (2008). Enriching ontological user profiles with tagging history for multi-domain recommendations.

Chen, L., & Pu, P. (2010). Eye-tracking study of user behavior in recommender interfaces. In *International conference on user modeling, adaptation, and personalization* (pp. 375–380).

Choroś, K. (2009). 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* (Vol. 64, p. 73-80). Springer Berlin Heidelberg.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011). Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Human-computer interaction–interact 2011* (pp. 152–168). Springer.

Cremonesi, P., Garzottto, F., & Turrin, R. (2012). User effort vs. accuracy in rating-based elicitation. In *Proceedings of the sixth acm conference on recommender systems* (pp. 27–34).

Cremonesi, P., Koren, Y., & Turrin, R. (2010). 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* (p. 39-46).

Dasiopoulou, S., Saathoff, C., Mylonas, P., Avrithis, Y., Kompatsiaris, Y., Staab, S., & Strintzis, M. (2008). *Semantic multimedia and ontologies theory and applications, chapter introducing context and reasoning in visual content analysis: an ontology-based framework*. Springer.

Datasets | grouplens. (n.d.). `http://grouplens.org/datasets/`. Accessed: 2015-05-01.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., … others (2010). The youtube video recommendation system. In *Proceedings of the fourth acm conference on recommender systems* (pp. 293–296).

de Gemmis, M., Lops, P., Musto, C., Narducci, F., & Semeraro, G. (2015). Semantics-aware content-based recommender systems. In *Recommender systems handbook* (pp. 119–159). Springer. doi: 10.1007/978-1-4899-7637-6_4

Degemmis, M., Lops, P., & Semeraro, G. (2007). A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, *17*(3), 217–255.

De Gemmis, M., Lops, P., Semeraro, G., & Basile, P. (2008). Integrating tags in a semantic content-based recommender. In *Proceedings of the 2008 acm conference on recommender systems* (pp. 163–170).

Deldjoo, Y., & Atani, R. E. (2016). A low-cost infrared-optical head tracking solution for virtual 3d audio environment using the nintendo wii-remote. *Entertainment Computing*, *12*, 9–27.

Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., & Piazzolla, P. (2016). Recommending movies based on mise-en-scene design. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 1540–1547).

Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., & Quadrana, M. (2016). Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, 1–15.

Deldjoo, Y., Elahi, M., Cremonesi, P., Moghaddam, F. B., & Caielli, A. L. E. (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).

Deldjoo, Y., Elahi, M., Quadrana, M., & Cremonesi, P. (2015). Toward building a content-based video recommendation system based on low-level features. In *E-commerce and web technologies*. Springer.

Deldjoo, Y., Elahi, M., Quadrana, M., & Cremonesi, P. (2018). Using visual features based on mpeg-7 and deep learning for movie recommendation. *International Journal of Multimedia Information Retrieval*.

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* (pp. 162–165).

Deldjoo, Y., Elahi, Y., Cremonesi, P., Moghaddam, F. B., & Caielli, A. L. E. (2017). How to combine visual features with tags to improve movie recommendation accuracy? In *E-commerce and web technologies: 17th international conference, ec-web 2016, porto, portugal, september 5-8, 2016, revised selected papers* (Vol. 278, p. 34).

Deldjoo, Y., Quadrana, M., Elahi, M., & Cremonesi, P. (2017). Using mise-en-sc\ene visual features based on mpeg-7 and deep learning for movie recommendation. *arXiv preprint arXiv:1704.06109*.

Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 143–177.

Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., & Zanker, M. (2012). Linked open data to support content-based recommender systems. In *Proceedings of the 8th international conference on semantic systems* (pp. 1–8).

Dorai, C., & Venkatesh, S. (2001). Computational media aesthetics: Finding meaning beautiful. *IEEE MultiMedia*, *8*(4), 10–12.

Dwivedi, P., & Bharadwaj, K. K. (2015). e-learning recommender system for a group of learners based on the unified learner profile approach. *Expert Systems*, *32*(2), 264–276.

Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). 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).

Elahi, M., Deldjoo, Y., Bakhshandegan Moghaddam, F., Cella, L., Cereda, S., & Cremonesi, P. (2017). Exploring the semantic gap for movie recommendations. In *Proceedings of the eleventh acm conference on recommender systems* (pp. 326–330).

Elahi, M., Ricci, F., & Rubens, N. (2013). Active learning strategies for rating elicitation in collaborative filtering: a system-wide perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 13.

Elahi, M., Ricci, F., & Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*.

Fleischman, M., & Hovy, E. (2003). Recommendations without user preferences: a natural language processing approach. In *Proceedings of the 8th international conference on intelligent user interfaces* (pp. 242–244).

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, *15*(1), 1–10.

Gedikli, F., & Jannach, D. (2013). Improving recommendation accuracy based on item-specific tag preferences. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(1), 11.

Georgiev, K., & Nakov, P. (2013). A non-iid framework for collaborative filtering with restricted boltzmann machines. In *International conference on machine learning* (pp. 1148–1156).

Guyon, I., Matic, N., Vapnik, V., et al. (1996). *Discovering informative patterns and data cleaning*.

Haghighat, M., Abdel-Mottaleb, M., & Alhalabi, W. (2016). Fully automatic face normalization and single sample face recognition in unconstrained environments. *Expert Systems with Applications*, *47*, 23–34.

Haley, G. M., & Manjunath, B. (1999). Rotation-invariant texture classification using a complete space-frequency model. *IEEE transactions on Image Processing*, *8*(2), 255–269.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, *16*(12), 2639–2664.

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *5*(4), 19.

Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *5*(4), 19.

Haskell, B. G., Puri, A., & Netravali, A. N. (1996). *Digital video: an introduction to mpeg-2*. Springer Science & Business Media.

Hawashin, B., Lafi, M., Kanan, T., & Mansour, A. (2019). An efficient hybrid similarity measure based on user interests for recommender systems. *Expert Systems*, e12471.

He, R., & McAuley, J. (2015). Vbpr: Visual bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1510.01784*.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hong, L., Doumith, A. S., & Davison, B. D. (2013). Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth acm international conference on web search and data mining* (pp. 557–566). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2433396.2433467` doi: 10.1145/2433396.2433467

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. In *1981 technical symposium east* (pp. 319–331).

Hornick, M. F., & Tamayo, P. (2012). Extending recommender systems for disjoint user/item sets: The conference recommendation problem. *IEEE Transactions on Knowledge and Data Engineering*(8), 1478–1490.

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). 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.

Insights, T. (2018). *500 hours of video uploaded to youtube every minute [forecast]*. `http://tubularinsights.com/hours-minute-uploaded-youtube/`.

Jakob, N., Weber, S. H., Müller, M. C., & Gurevych, I. (2009). Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In *Proceedings of the 1st international cikm workshop on topic-sentiment analysis for mass opinion* (pp. 57–64).

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction*. Cambridge University Press.

Karimi, R., Nanopoulos, A., & Schmidt-Thieme, L. (2015). A supervised active learning framework for recommender systems based on decision trees. *User Modeling and User-Adapted Interaction*, *25*(1), 39–64.

Kasutani, E., & Yamada, A. (2001). The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Image processing, 2001. proceedings. 2001 international conference on* (Vol. 1, pp. 674–677).

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. In *Acm sigir forum* (Vol. 37, pp. 18–28).

Khamparia, A., & Singh, K. M. (2019). A systematic review on deep learning architectures and applications. *Expert Systems*, *36*(3), e12400.

Knees, P., Pohle, T., Schedl, M., & Widmer, G. (2007). 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).

Kohavi, R. (1995). The power of decision tables. In *8th european conference on machine learning* (p. 174-189). Springer.

Kohavi, R., & Sommerfield, D. (1998). Targeting business users with decision table classifiers. In *Kdd* (pp. 249–253).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 ieee computer society conference on computer vision and pattern recognition (cvpr'06)* (Vol. 2, pp. 2169–2178).

Lehinevych, T., Kokkinis-Ntrenis, N., Siantikos, G., Dogruöz, A. S., Giannakopoulos, T., & Konstantopoulos, S. (2014). 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).

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *2*(1), 1–19.

Li, S., Cheng, X., Su, S., & Sun, H. (2017). Exploiting organizer influence and geographical preference for new event recommendation. *Expert Systems*, *34*(2), e12190.

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

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.

Lops, P., De Gemmis, M., Semeraro, G., Musto, C., & Narducci, F. (2013). Content-based and collaborative techniques for tag recommendation: an empirical evaluation. *Journal of Intelligent Information Systems*, *40*(1), 41–61.

Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., & Hellerstein, J. M. (2012). Distributed graphlab: a framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment*, *5*(8), 716–727.

Ma, W.-Y., & Manjunath, B. (1998). A texture thesaurus for browsing large aerial photographs. *Journal of the American Society for Information Science*, *49*(7), 633–648.

Magnini, B., & Strapparava, C. (2001). Improving user modelling with content-based techniques. In *User modeling 2001* (pp. 74–83). Springer.

Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., & Yamada, A. (2001). Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, *11*(6), 703–715.

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). *Introduction to mpeg-7: multimedia content description interface* (Vol. 1). John Wiley & Sons.

Martins, E. F., Belém, F. M., Almeida, J. M., & Gonçalves, M. A. (2016). On cold start for associative tag recommendation. *Journal of the Association for Information Science and Technology*, *67*(1), 83–105.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 54–88.

Milicevic, A. K., Nanopoulos, A., & Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, *33*(3), 187–209.

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth acm conference on digital libraries* (pp. 195–204).

Musto, C., Narducci, F., Lops, P., Semeraro, G., de Gemmis, M., Barbieri, M., … Clout, R. (2012). Enhanced semantic tv-show representation for personalized electronic program guides. In *User modeling, adaptation, and personalization* (pp. 188–199). Springer.

Nakache, D., Metais, E., & Timsit, J. F. (2005). Evaluation and nlp. In *Database and expert systems applications* (pp. 626–632).

Ning, X., Desrosiers, C., & Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook* (pp. 37–76). Springer.

Ning, X., & Karypis, G. (2012). Sparse linear methods with side information for top-n recommendations. In *Proceedings of the sixth acm conference on recommender systems* (pp. 155–162).

Ohm, J.-R., Kim, H., & Krishnamachari, S. (2005). The mpeg-7 color descriptors..

Pazzani, M. J., & Billsus, D. (2007a). The adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), (pp. 325–341). Berlin, Heidelberg: Springer-Verlag. Retrieved from `http://dl.acm.org/citation.cfm?id=1768197.1768209`

Pazzani, M. J., & Billsus, D. (2007b). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Springer.

Rasheed, Z., & Shah, M. (2003). Video categorization using semantics and semiotics. In *Video mining* (pp. 185–217). Springer.

Rasheed, Z., Sheikh, Y., & Shah, M. (2005). On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, *15*(1), 52–64.

Renckes, S., Polat, H., & Oysal, Y. (2012). A new hybrid recommendation algorithm with privacy. *Expert Systems*, *29*(1), 39–55.

Rendle, S. (2010). Factorization machines. In *Data mining (icdm), 2010 ieee 10th international conference on* (pp. 995–1000).

Rendle, S. (2012a). Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(3), 57.

Rendle, S. (2012b, May). Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol.*, *3*(3), 57:1–57:22. Retrieved from `http://doi.acm.org/10.1145/2168752.2168771` doi: 10.1145/2168752.2168771

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (pp. 452–461).

Rendle, S., Gantner, Z., Freudenthaler, C., & Schmidt-Thieme, L. (2011). Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 635–644).

Resnick, P., & Varian, H. R. (1997a). Recommender systems. *Communications of the ACM*, *40*(3), 56–58.

Resnick, P., & Varian, H. R. (1997b). Recommender systems. *Commun. ACM*, *40*(3), 56–58. doi: http://doi.acm.org/10.1145/245108.245121

Resnick, P., & Varian, H. R. (1997c). Recommender systems. *Commun. ACM*, *40*(3), 56–58. doi: http://doi.acm.org/10.1145/245108.245121

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (p. 1-35). Springer Verlag.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Springer US.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender systems handbook*. Springer.

Rubens, N., Elahi, M., Sugiyama, M., & Kaplan, D. (2015). Active learning in recommender systems. In *Recommender systems handbook* (pp. 809–846). Springer.

Sainath, T. N., Kingsbury, B., Sindhwani, V., Arisoy, E., & Ramabhadran, B. (2013). Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on* (pp. 6655–6659).

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on machine learning* (pp. 791–798).

Santos, O. C., & Boticario, J. G. (2015). User-centred design and educational data mining support during the recommendations elicitation process in social online learning environments. *Expert Systems*, *32*(2), 293–311.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data mining and knowledge discovery*, *5*(1-2), 115–153.

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & Elahi, M. (2017). Current challenges and visions in music recommender systems research. *arXiv preprint arXiv:1710.03208*.

Seyerlehner, K., Schedl, M., Pohle, T., & Knees, P. (2010). Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*.

Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 acm conference on recommender systems* (pp. 259–266).

Shi, Y., Larson, M., & Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, *47*(1), 3.

Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., & Trancoso, I. (2011). Temporal video segmentation to scenes using high-level audiovisual features. *Circuits and Systems for Video Technology, IEEE Transactions on*, *21*(8), 1163–1177.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, *2009*, 4:2–4:2.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., & Servedio, V. D. (2007). Folksonomies, the semantic web, and movie recommendation.

Tkalčič, M., Maleki, N., Pesek, M., Elahi, M., Ricci, F., & Marolt, M. (2017). A research tool for user preferences elicitation with facial expressions. In *Proceedings of the eleventh acm conference on recommender systems* (pp. 353–354).

Tkalcic, M., & Tasic, J. F. (2003). Colour spaces: perceptual, historical and applicational background. In *Eurocon 2003. computer as a tool. the ieee region 8* (Vol. 1, pp. 304–308).

TURI. (2018). *Graphlab recommender factorization recommender.* Retrieved from `https://turi.com/products/create/docs/generated/graphlab.recommender.factorization_recommender.FactorizationRecommender.html`

Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. *Journal of Experimental Psychology: General*, *123*(4), 394.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems* (pp. 2643–2651).

Vargas, S., & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth acm conference on recommender systems* (pp. 109–116).

Vig, J., Sen, S., & Riedl, J. (2009). Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on intelligent user interfaces* (pp. 47–56).

Vinagre, J., Jorge, A. M., & Gama, J. (2018). Online bagging for recommender systems. *Expert Systems*, *35*(4), e12303.

Vlachos, M., Duenner, C., Heckel, R., Vassiliadis, V. G., Parnell, T., & Atasu, K. (2018). Addressing interpretability and cold-start in matrix factorization for recommender systems. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, H. L., & Cheong, L.-F. (2006). Affective understanding in film. *Circuits and Systems for Video Technology, IEEE Transactions on*, *16*(6), 689-704.

Wang, L., Zeng, X., Koehl, L., & Chen, Y. (2015). Intelligent fashion recommender system: Fuzzy logic in personalized garment design. *IEEE Trans. Human-Machine Systems*, *45*(1), 95–109.

Wang, X.-Y., Zhang, B.-B., & Yang, H.-Y. (2014). Content-based image retrieval by integrating color and texture features. *Multimedia tools and applications*, *68*(3), 545–569.

Wang, Y., Xing, C., & Zhou, L. (2006). Video semantic models: survey and evaluation. *Int. J. Comput. Sci. Netw. Security*, *6*, 10–20.

Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., & Li, M. (2007). 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).

Youtube. (n.d.). `http://www.youtube.com`. Accessed: 2015-04-01.

Zabih, R., Miller, J., & Mai, K. (1996). Video browsing using edges and motion. In *Computer vision and pattern recognition, 1996. proceedings cvpr'96, 1996 ieee computer society conference on* (pp. 439–446).

Zettl, H. (2002). Essentials of applied media aesthetics. In C. Dorai & S. Venkatesh (Eds.), *Media computing* (Vol. 4, p. 11-38). Springer US.

Zettl, H. (2013). *Sight, sound, motion: Applied media aesthetics*. Cengage Learning.

Zhao, X., Li, G., Wang, M., Yuan, J., Zha, Z.-J., Li, Z., & Chua, T.-S. (2011). Integrating rich information for video recommendation with multi-task rank aggregation. In *Proceedings of the 19th acm international conference on multimedia* (pp. 1521–1524).

Zhou, H., Hermans, T., Karandikar, A. V., & Rehg, J. M. (2010). Movie genre classification via scene categorization. In *Proceedings of the international conference on multimedia* (pp. 747–750).

AUTHOR BIOGRAPHY

Naieme Hazrati is a PhD student at Free University of Bozen - Bolzano (Italy). She has received M.Sc. in Computer Science from University of Tehran (Iran). Her research focuses include recommender systems, social network analysis, and content-based recommender systems.

Mehdi Elahi. Mehdi Elahi is an associate professor at University of Bergen (Norway). He received M.Sc. degree in Electrical Engineering (Sweden) in 2010, and Ph.D. degree in Computer Science (Italy) in 2014. Over the last 3 years, he has been serving as an assistant professor at Free University of Bozen - Bolzano (Italy), where has researched on various aspects of Recommender Systems. As a result of his research work, he served as a primary author or co-author of more than 60 peer-reviewed publications in AI, RS, and HCI related conferences and journals. He has been actively involved in co-authorship of a US-patent as well as co-authorship of several EU research proposals. He has been awarded a number of industry and academic research grants, e.g., by a world-class company (Amazon), and a well-known academic institute in Italy (Polytechnic University of Milan). He has been actively involved in research and development of up-and-running mobile recommender systems for food (ChefPad) and for tourism domain (South Tyrol Suggests). He has provided various types of community services such as co-organization of the ACM RecSys challenge 2017 (organized by XING), and advisor of RecSys challenge 2018 (organized by Spotify).

How to cite this article: Naieme Hazrati, Mehdi Elahi, (2016), Addressing the New Item Problem in Video Recommender Systems by Incorporation of Visual Features with Restricted Boltzmann Machines, *Q.J.R. Meteorol. Soc., 2017;00:1–6.*