IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 21, 2019, accepted April 8, 2019, date of publication April 11, 2019, date of current version April 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910722

Exploiting Aesthetic Features in Visual Contents for Movie Recommendation

XIAOJIE CHEN[1], PENGPENG ZHAO[1,2], YANCHI LIU[3], LEI ZHAO[1], JUNHUA FANG[1], VICTOR S. SHENG[4], AND ZHIMING CUI[5]

[1]Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3]Management Science and Information Systems, Rutgers University, New Brunswick, NJ 08901-8554, USA
[4]Computer Science Department, University of Central Arkansas, Conway, AR 72035, USA
[5]School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

Corresponding author: Pengpeng Zhao (ppzhao@suda.edu.cn)

This work was supported in part by the NSFC under Grant 61876217, Grant 61872258, and Grant 61728205, in part by the Suzhou Science and Technology Development Program under Grant SYG201803, in part by the Postdoctoral Research Foundation of China under Grant 2017M621813, in part by the Natural Science Fund for Colleges and Universities in Jiangsu Province under Grant 18KJB520044, and in part by the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, under Grant IIP2019-1.

ABSTRACT As one of the most widely used recommender systems, movie recommendation plays an important role in our life. However, the data sparsity problem severely hinders the effectiveness of personalized movie recommendation, which requires more rich content information to be utilized. Posters and still frames, which directly display the visual contents of movies, have significant influences on movie recommendation. They not only reveal rich knowledge for understanding movies but also useful for understanding user preferences. However, existing recommendation methods rarely consider aesthetic features, which tell how the movie looks and feels, extracted from these pictures for the movie recommendation. To this end, in this paper, we propose an aesthetic-aware unified visual content matrix factorization (called UVMF-AES) to integrate visual feature learning and recommendation into a unified framework. Specifically, we first integrate the convolutional neural network (CNN) features and aesthetic features into probabilistic matrix factorization. Then we establish a unified optimization framework with these features for the movie recommendation. The experimental results on two real-world datasets show that our proposed method UVMF-AES is significantly superior to the state-of-the-art methods on movie recommendation.

INDEX TERMS Movie recommendation, aesthetic features, probabilistic matrix factorization.

I. INTRODUCTION

With the development of information technology and rapid data accumulation, recommender systems have become a key part of our daily life for fetching useful information. Recent years have witnessed a revolution in recommender systems from both methodology and application perspectives. As one of the most widely used recommender systems, movie recommendation is a subtle way of influencing our lives. With the comprehensive analysis of users' historical behaviors, user preferences, social relations, etc., the goal of a movie recommendation system is to recommend movies to potentially interested users. However, the high sparsity of user-item interactions, which is common in movie recommendation scenario where the user and movie sets are extremely large, severely hinders the effectiveness of personalized recommendation.

Existing studies have shown that content-based recommendation using rich information can alleviate this problem [1]–[4]. For example, content-based movie recommendations using content information including movie properties, user demographics, movie reviews, etc., have been successfully applied. Figure 1 shows posters and still frames for two movies respectively. From this figure, we can infer that users who delight in movie 1 may also be fond of movie 2 because they have similar posters and still frames. On one hand, we can infer from the two similar frames that both men are riding motorcycle, which is depicted by the pictures. On the other hand, both frames, which look wonderful and make audiences feel good, have high-level attributes in aesthetic assessment. In recent years, several researches [5]–[8] have considered visual contents for recommendation, such as photos, posters, and key frames. From these images we

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

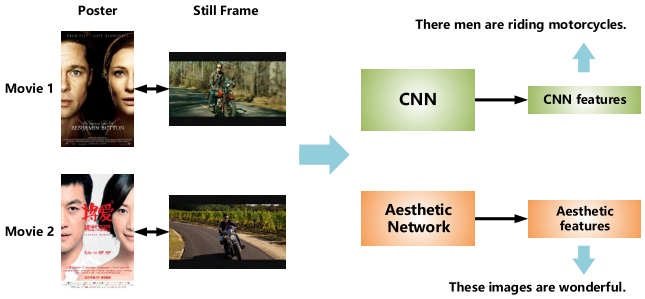

FIGURE 1. CNN features and Aesthetic features of movie posters and still frames.

are able to extract two kinds of features, i.e., Convolutional Neural Network (CNN) feature and aesthetic feature. As a part of visual contents, CNN features have been gradually taken into consideration. However, as an essential component of content, aesthetic features, which reveal rich knowledge for understanding movies and user preferences, are rarely considered.

In this paper, we focus on how to utilize aesthetic features together with CNN features of images to improve the performance of movie recommendation. To this end, we propose a model named Aesthetic-aware Unified Visual contents Matrix Factorization (UVMF-AES) to exploit aesthetic features in movie posters and still frames for recommendation. Moreover, UVMF-AES integrates aesthetic feature extraction and recommendation into a unified framework which can be trained end-to-end. Specifically, we integrate aesthetic features and CNN features into probabilistic matrix factorization (PMF), so that visual features can be optimized for movie recommendation. The contributions of this work can be summarized as follows:

- We propose to enhance movie recommendation with aesthetic features extracted from movie posters and still frames in addition to CNN features.
- We develop an Aesthetic-aware Unified Visual contents Matrix Factorization model (UVMF-AES), which integrates aesthetic feature extraction and recommendation into a unified framework.
- Our experimental results on real-world datasets reveal that our proposed method UVMF-AES is capable of leveraging aesthetic features effectively.

II. RELATED WORK

This paper exploiting aesthetic features for movie recommendation. Starting from content-based recommendation, we incorporate CNN features and aesthetic features extracted from deep learning networks into probabilistic matrix factorization. Hence, we review related work on collaborative filtering, content-based recommendation, visual contents for recommendation, and aesthetic assessment.

A. COLLABORATIVE FILTERING AND CONTENT-BASED RECOMMENDATION

Collaborative filtering is a widespread method and lays the foundation of present recommender system. As we know, collaborative filtering methods pay attention to the relationship between users and items. Jeong *et al.* [9] utilize collaborative filtering to develop a movie recommender system. The similarity between each movie is determined by the similarity of ratings of these movies by the users who rated both movies. Matrix factorization [10]–[13] is one of widely used collaborative filtering method. Unlike collaborative filtering, content-based recommendation methods pay attention to properties of items. The similarity of items' properties decides the similarity between each item. So content-based methods help to alleviate data sparsity and cold start problem. Christakou *et al.* [14] integrate a content-based method with a collaborative filtering method and develop a recommender system. Salter and Antonopoulos [15] develop a recommender agent which both use collaborative filtering and content-based method. With the development of deep learning, many excellent research works use deep learning network to train the content-based recommendation model. To evaluate music recommender system, Van den Oord *et al.* [16] construct a reasonable ground truth dataset on the basis of the content of personal music databases, and show good results for a novel content-based recommendation that uses a mixture of CNN features and audio features than a standard low-level MFCC(mel-frequency cepstral coefficients) feature representation. Traditional content-based music recommendation cannot extract the representative audio content features, which can contain important information. Wang and Wang [17] present a deep learning model to integrate automatic learning features and collaborative filtering.

B. VISUAL CONTENTS FOR RECOMMENDATION

As an image can contain much information, it may help to alleviate data sparsity problem in recommendation or other data mining tasks. He *et al.* [18] proposed a hierarchical embedding architecture which accounts for both high-level and subtle visual features. As fashion gradually becomes the focus of the masses, McAuley *et al.* [19] propose a scalable method to explain the visual relationship of human sense. Aimed at recommending clothing matching scheme, they define the problem and provide the dataset for training. He and McAuley [20] integrate high-level visual features, and user feedback before and dynamic trends to build a novel model for one-class collaborative filtering. Kalantidis *et al.* [21] provide a scalable method to recommend clothing to users automatically. They mainly check the existing categories in the clothing. Essentially they discover the potential clothing by measuring the similarity between the clothing. Instagram provides the function that photos relevant to POIs can be shared online and for users. Li *et al.* [22] propose a personalized ranking method to integrate visual and textual information for content-based personalized recommendation. Zhao *et al.* [8] mainly focus on the similarity of visual features. The visual features are extracted from the photos taken by users at each POI. According to the user's interests which are analyzed by matrix factorization, the goal is to recommend tourist routes to travelers. Chen *et al.* [5] take visual contents into recommendation on the basis of

matrix factorization. Around content-based recommendation, CNN features extracted from posters and still frames are used as a supplement to the content.

C. AESTHETIC ASSESSMENT

With the wide application of deep learning network, more and more aesthetic networks are used for aesthetic assessment. AVA [23], a well-known large-scale aesthetic analysis database, merges and pushes forward the aesthetic networks. Yu *et al.* [24] extracted aesthetic features by a pre-trained neural network, and proposed a new tensor factorization model to incorporate the aesthetic features in a clothing recommendation task. Mai *et al.* [25] proposed a MNA-CNN network, which can extract features at multiple scales and naturally incorporates scene categories for aesthetic assessment. Zhou *et al.* [26] and Talebi and Milanfar [27] proposed a CNN-based image assessment method, which can predict the distribution of quality ratings. Lv *et al.* [28] first proposed a method to construct an exclusive dataset to represent the aesthetic preference of each individual. Furthermore, they proposed a customized aesthetic distribution model based on the dataset constructed well-designed aesthetic attributes in the first step. Finally, they concluded a user-friendly aesthetic ranking framework via deep neural network. Deng *et al.* [29] proposed an EnhanceGAN model, which can leverage abundant images and permit multiple forms of color enhancement. Their model is extensible to include further image enhancement schemes. Sheng *et al.* [30] developed an end-to-end way to train an assessment model with aesthetic labels only. They presented three typical attention mechanisms, including average, minimum, and adaptive.

Inspired by related work, we can find that visual contents, such as CNN features and aesthetic features, do help to enhance the performance of movie recommendation. Moreover, we focus on how to integrate CNN features and aesthetic features into recommendation.

III. PRELIMINARIES

In this section, we first introduce problem definition of the movie recommendation problem. We next introduce the basic recommendation model, probabilistic matrix factorization.

A. PROBLEM DEFINITION

There are three objects, namely, users, movies, and images in our movie recommendation problem. Let $\mathbb{U} = \{u_1, u_2, \ldots, u_i, \ldots, u_N\}$ be the set of N users and $\mathbb{V} = \{v_1, v_2, \ldots, v_j, \ldots, v_M\}$ be the set of M movies. As each user has watching records, there are ratings that users rated the movies they watched. R denotes the observed user-movie rating matrix, and $R \in \mathrm{R}^{N\times M}$. We use R_{ij} to denote the ratings of a user u_i on a movie v_j. $U \in \mathrm{R}^{K\times N}$ denotes the user feature matrix, and $V \in \mathrm{R}^{K\times M}$ denotes the movie feature matrix. u_i denotes the user latent feature of u_i, and v_j denotes the movie latent feature of v_j. Here, images mainly include movie posters and still frames. As parts of movies' properties, each movie v_j has its own posters and still frames. So all the images can be collected in the set $\mathbb{P} = \{p_1, p_2, \ldots, p_l, \ldots, p_L\}$, where L is the total number of these images. One set $\mathbb{P}_{v_j}$ can be used to denote the posters and the still frames which belong to movie v_j. Every set of movie's images can be collected in the set $\mathbb{P}$, and the relationship can be described as $\mathbb{P} = \mathbb{P}_{v_1} \cup \mathbb{P}_{v_2} \cup \ldots \cup \mathbb{P}_{v_j} \cup \ldots \cup \mathbb{P}_{v_M}, j = 1, \ldots, M$. The problem can be described as follows: Given N users and M movies, and images of movies $\mathbb{P}$, accroding to the prediction scores, the goal of our movie recommendation problem is to recommend top-N movies to each user.

B. A BASIC RECOMMENDATION MODEL

Probabilistic Matrix Factorization(PMF) is based on the following two hypotheses; more details can be found in [31]. First, PMF defines Gaussian distribution over the difference between the observed rating matrix and the predicted rating matrix. So the conditional distribution over the observed ratings can be estimated as follows.

$$P(R|U, V, \sigma) = \prod_{i=1}^{N}\prod_{j=1}^{M}(\mathcal{N}(R_{ij}|u_i^T v_j, \sigma^2))^{Y_{ij}} \tag{1}$$

where column vectors u_i and v_j are used to describe user-specific and movie-specific latent feature vectors, respectively. $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of Gaussian distribution with mean μ and variance σ^2. Y signifies the indicator function, if $R_{ij} > 0$, $Y_{ij} = 1$. Otherwise, $Y_{ij} = 0$.

Second, PMF defines Gaussian distribution over the latent matrices U and V, $P(U|\sigma_u^2) = \prod_{i=1}^{N}\mathcal{N}(u_i|0, {\sigma_u}^2 I)$ and $P(V|\sigma_v^2) = \prod_{j=1}^{M}\mathcal{N}(v_j|0, {\sigma_v}^2 I)$. According to the hypotheses, the posterior distribution can be obtained as follows.

$$P(U,V|R) = \prod_{i=1}^{N}\mathcal{N}(u_i|0, {\sigma_u}^2 I) \cdot \prod_{j=1}^{M}\mathcal{N}(v_j|0, {\sigma_v}^2 I) \cdot \prod_{i=1}^{N}\prod_{j=1}^{M}(\mathcal{N}(R_{ij}|u_i^T v_j, \sigma^2))^{Y_{ij}} \tag{2}$$

IV. MOVIE RECOMMENDATION WITH CNN FEATURES AND AESTHETIC FEATURES

A. EXTRACTING CNN FEATURES AND AESTHETIC FEATURES

In this step, visual contents in our studied problem include CNN features and aesthetic features. Here, we use VGG16 [32] model to extract CNN features. CNN features focus on the content of image, and mainly explain what the image is. VGG16 model has low error rate in classification and strong extensiveness, so VGG16 is widely used to image feature extraction. From Figure 2, by repeatedly connecting 3×3 convolution kernels and 2×2 pooling kernels, VGG16 increases the depth of the network structure to improve performance. Moreover, the dimension of the last fully connected layer and the softmax layer is 1000. Indeed, the dimension of the last two layers is the number of classification. For the 1×1000 column vector, the value in each

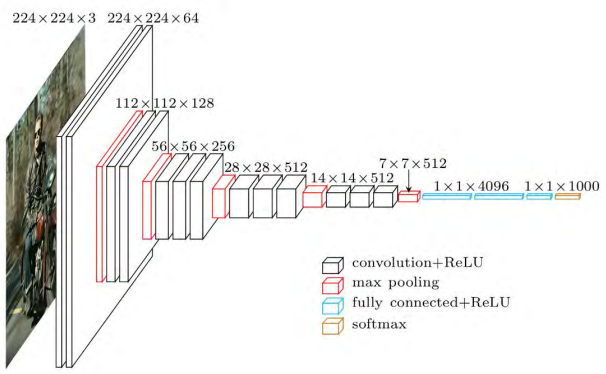

FIGURE 2. The graphical architecture of VGG16.

dimension means the probability of each category. Therefore, the feature vector cannot reflect the content of image. As a result, the last two layers are designed for classification. So we remove the last two layers. Given an input image of size 224×224×3, where 224×224 is the size of image, 3 is the number of RGB channels, the dimension D of CNN features is 4096. In the beginning, VGG16 is initiated with pre-trained weights on Imagenet [33], and then the CNN features can be fine-tuned [34].

Then we use a deep aesthetic framework OWACNN [35] to extract the aesthetic features. From Figure 3, we can find that it contains the following operations. First, the input is an image of size 224×224×3, where 224×224 is the size of an image, 3 is the number of RGB channels. Second, we use a pre-trained ResNet-152 [36] as the expert network. The expert network is made up of four different pre-trained models, scene recognition, semantic analysis, object recognition, and emotion prediction. Third, it connects an OWA pooling layer, which is an aggregation operator. Then, the feature is aggregated and the dimension of features is 4096. Finally, the framework uses a full-connected layer to make an AQ (Aesthetic Quality) decision. As the full-connected layer is designed for classification, we remove the last full-connected layer. After extracting the visual features from CNN, the following step is to model these features into recommendation.

B. MODELING CNN FEATURES AND AESTHETIC FEATURES

Suppose that we have an image p_k relevant with a movie v_j, we can use the visual content of p_k to describe movie v_j. An arbitrary image p_s, which is not relevant with movie v_j, may have a smaller probability to be used to describe movie v_j. Here, visual content includes CNN features and aesthetic features. We use *CNN* to denote the CNN features of an image and *AES* to denote the aesthetic features of an image. Item features v_j can be used to describe movie v_j, so the latent feature v_j implies whether the image p_s is relevant with movie v_j or not. Thus, we use $P(f_{jk} = 1|v_j, p_k)$ to describe p_k is relevant with v_j,

$$P(f_{jk} = 1|v_j, p_k) = \frac{exp(v_j^T \cdot W \cdot f_{p_k})}{\sum_{l=1}^{L} exp(v_j^T \cdot W \cdot f_{p_l})} \tag{3}$$

where $W \in \mathrm{R}^{K \times D}$ is the interaction matrix between visual features and item features, and K and D are the dimension of item features and visual features, respectively. $f_{p_k} = \begin{bmatrix} CNN_{p_k} \\ AES_{p_k} \end{bmatrix}$ is the concatenation of CNN features(*CNN*) and aesthetic features(*AES*). Thus, for $p_k \in \mathbb{P}_{v_j}$, by maximizing $P(f_{jk} = 1|v_j, p_k)$, we tie v_j and f_{p_k} closely together through W. In this way, visual content is integrated into the learning process of v_j.

$$P(\mathbb{F}|\mathbb{P}, U, V, W) = \prod_{j=1}^{M} \prod_{p_k \in \mathbb{P}_{v_j}} P(f_{jk} = 1|v_j, p_k) \tag{4}$$

where $\mathbb{F} = \{f_{jk} : p_k \in \mathbb{P}_{v_j}, \forall v_j \in \mathbb{V}\}$. We assume Gaussian prior for W as $P(W|\sigma_w) = \prod_{i=1}^{K} \prod_{j=1}^{D} \mathcal{N}(W_{ij}|0, \sigma_w^2)$, where σ_w^2 are the variances.

With incorporating the rating data and the visual features, we propose an Aesthetic-aware Unified Visual Contents Matrix Factorization model (called UVMF-AES). We use $\mathcal{F}$ to denote the following objective function,

$$\max_{U,V,W,f} \log P(U, V, W|R, \mathbb{F}, \mathbb{P}) \tag{5}$$

where the posterior distribution $P(U, V, W|R, \mathbb{F}, \mathbb{P})$ can be as follows,

$$\begin{aligned} &P(U, V, W|R, \mathbb{F}, \mathbb{P}) \\ &\quad \propto P(R, \mathbb{F}|U, V, W, \mathbb{P})P(U, V, W|\mathbb{P}) \\ &\quad = P(R|U, V)P(\mathbb{F}|U, V, W)P(W)P(U)P(V) \end{aligned} \tag{6}$$

Figure 4 shows the graphic description of our proposed model UVMF-AES. Its right is a PMF part (green-dotted), and its left is a CNN part (red-dashed). As a part of movies' properties, visual content can be used to reflect the properties of movies. In addition, UVMF-AES defines Gaussian distribution over the weights W of both CNN and aesthetic network (OWACNN). UVMF-AES takes advantage of PMF to integrate *CNN* and *AES* into the recommendation.

By replacing Eq.(2), Eq.(4), Eq.(6), the objective function $\mathcal{F}$ in Eq.(5) is as follows,

$$\begin{aligned} \max_{U,V,W,f} &- \left\| Y \odot (R - U^T V) \right\|_F^2 - \lambda_1(\|U\|_F^2 + \|V\|_F^2) \\ &+\alpha \sum_{j=1}^{M} \sum_{p_k \in \mathbb{P}_{v_j}} \log P(f_{jk} = 1|v_j, p_k) - \lambda_2 \|W\|_F^2 \end{aligned} \tag{7}$$

where we set $\lambda_1 = \frac{\sigma^2}{\sigma_u^2} = \frac{\sigma^2}{\sigma_v^2}$, $\lambda_2 = \frac{\sigma^2}{\sigma_w^2}$ for preventing over-fitting and $\alpha = \sigma^2$. $\odot$ is the entrywise product, $\|\cdot\|_F$ is Frobenius norm.

V. OPTIMIZATION

In this section, we adopt negative sampling and gradient ascent for optimization.

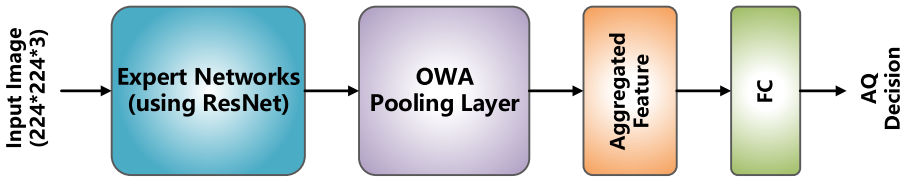

FIGURE 3. The macro-architecture of OWACNN framework.

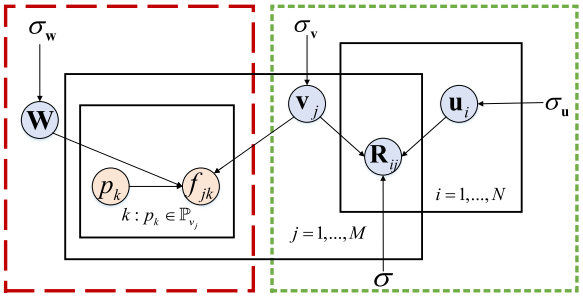

FIGURE 4. Graphical model of UVMF-AES.

A. NEGATIVE SAMPLING

Since there are so many images in the dataset, it takes too much time in the convolutional operation. The formula of the objective function involves the calculation of visual features. We adopt negative sampling [37] to accelerate the training speed. Approximately $\log P(f_{jk} = 1|v_j, p_k)$ are as follows,

$$\log \sigma(v_j^T \cdot W \cdot f_{p_k}) + \sum_{t=1}^{J} \log \sigma(-v_j^T \cdot W \cdot f_{p_{kt}}) \tag{8}$$

where p_{kt} $(t = 1, \ldots, J)$ are named negative samples, J is the number of negative samples for p_k. Given each training sample p_k, the relationship between p_k and its negative samples p_{kt} $(t = 1, \ldots, J)$ is one-to-many. For each image $p_k \in \mathbb{P}$, negative samples can be collected in $\mathbb{X}_{p_k} = \{p_{k1}, \ldots, p_{kt}, \ldots, p_{kJ}\}$.

Generally, point at each image p_k of v_j, we randomly select J images as negative samples, i.e. p_{kt}. Here, the selected images are not from the image set of v_j. The goal of optimization is to maximize the similarity between v_j and f_{p_k}, and minimize the similarity between v_j and $f_{p_{kt}}$. By negative sampling, the gradients can be simplified. And the updates rules can be as follows.

1) UPDATE U

The partial derivatives of $\mathcal{F}$ w.r.t U are given by,

$$\frac{\partial \mathcal{F}}{\partial U} = 2V(Y \odot R)^T - 2V[Y \odot (U^T V)]^T - 2\lambda_1 U \tag{9}$$

2) UPDATE V

The partial derivatives of $\mathcal{F}$ w.r.t V are given by,

$$\frac{\partial \mathcal{F}}{\partial V} = 2U(Y \odot R) - 2V[Y \odot (U^T V)] - 2\lambda_1 V + \alpha X \tag{10}$$

where X is the partial derivative of the term $\sum_{j=1}^{M} \sum_{p_k \in \mathbb{P}_{v_j}} \log P(f_{jk} = 1|v_j, p_k)$ in $\mathcal{F}$ w.r.t V. And $X = \{x_1, \ldots, x_j, \ldots, x_M\} \in \mathrm{R}^{K \times M}$ is a matrix with its j-th column x_j given as,

$$x_j = \sum_{p_k \in \mathbb{P}_{v_j}} [(1 - \sigma(v_j^T \cdot W \cdot f_{p_k})) \cdot W \cdot f_{p_k} - \sum_{t=1}^{J} (1 - \sigma(-v_j^T \cdot W \cdot f_{p_{kt}})) \cdot W \cdot f_{p_{kt}}] \tag{11}$$

Further, we can convert x_j into a form of vector representation,

$$x_j = W \cdot H_j(1_{|P_{v_j}|} - \sigma(H_j^T \cdot W^T) \cdot v_j) - W \cdot \widetilde{H}_j(1_{r \cdot |P_{v_j}|} - \sigma(\widetilde{H}_j^T \cdot W^T \cdot v_j)) \tag{12}$$

where the matrix $H_i \in \mathrm{R}^{D \times |P_{v_j}|}$ is made up of each column $f_{p_k}, p_k \in \mathbb{P}_{v_j}$. Similarly, the matrix $\widetilde{H}_j \in \mathrm{R}^{D \times r \cdot |P_{v_j}|}$ is made up of each column $f_{p_{kt}}$ where $p_k t$ is the negative samples of p_k.

3) UPDATE W

The partial derivatives of $\mathcal{F}$ w.r.t W are given by,

$$\frac{\partial \mathcal{F}}{\partial W} = \alpha \sum_{j=1}^{M} \sum_{p_k \in \mathbb{P}_{v_j}} [(1 - \sigma(v_j^T \cdot W \cdot f_{p_k})) \cdot v_j \cdot f_{p_k}^T - \sum_{t=1}^{J} (1 - \sigma(-v_j^T \cdot W \cdot f_{p_{kt}})) \cdot v_j \cdot f_{p_{kt}}^T] - 2\lambda_2 W \tag{13}$$

B. FINE-TUNING CNN

Each movie is usually done by an aesthetic director. We can just use OWACNN to extract the aesthetic features of each movie. To update the parameters for CNN, we fix U, V, W and remove irrelevant items, the partial derivative of $\mathcal{F}$ w.r.t θ is as follows,

$$\frac{\partial \mathcal{F}}{\partial \theta} = \sum_{j=1}^{M} \sum_{p_k \in \mathbb{P}_{v_j}} [(1 - \sigma(v_j^T \cdot W \cdot CNN(p_k))) \cdot \sum_{d=1}^{D} w_d^T v_j \frac{\partial CNN(p_k)_d}{\partial \theta} - \sum_{t=1}^{J} (1 - \sigma(-v_j^T \cdot W \cdot CNN(p_{kt}))) \sum_{d=1}^{D} w_d^T v_j \frac{\partial CNN(p_{kt})_d}{\partial \theta}] \tag{14}$$

where θ is the set of CNN weights to be tuned, and $CNN(p_k)_d$ denotes the d-th element of $CNN(p_k)$. We can use a back-propagation (BP) algorithm [38] to calculate the gradients of CNN. Here, $\frac{\partial \mathcal{F}}{\partial \theta}$ includes the gradients of CNN are involved in.

C. LEARNING ALGORITHM

The learning algorithm of UVMF-AES can be summarized in Algorithm 1. Given M movies and N users, their rating matrix R and images of movie $\mathbb{P}_{v_j}, j = 1, \ldots, M$, we aim to complete the recommendation task. In line 1 and 2, firstly we initialize the weights of VGG16 and U, V, W with Gaussian distribution respectively. From line 3 to line 8, the parameters are updated until convergence, where η is the learning rate. After training and optimizing, we get the predicted matrix of user-movie by $U^T V$. At last, we can rank the predicted values and select top-N movies for each user.

Algorithm 1 The Learning Algorithm of UVMF-AES

Require: R, $\mathbb{P}_{v_j}$ for $v_j \in \mathbb{V}$
Ensure: the full rating matrix of user-movie
1: Initialize VGG16, OWACNN by using pre-trained weights on ImageNet
2: Initialize U, V, W with Normal distribution $\mathcal{N}(0, 0.01)$
3: **repeat**
4: Update U as $U \leftarrow U + \eta \frac{\partial F}{\partial U}$
5: Update V as $V \leftarrow V + \eta \frac{\partial F}{\partial V}$
6: Update W as $W \leftarrow W + \eta \frac{\partial F}{\partial W}$
7: fine-tune *CNN* using back propagation and concatenate *AES*
8: **until** convergence;
9: **return** the full rating matrix of user-movie on $U^T V$

VI. EXPERIMENTS

A. DATASET

We evaluate the performance of our proposed method UVMF-AES on the MovieLens2011[1] and IMDb[2] datasets. For missing movie posters and still frames in the MovieLens2011 dataset, we crawled these image data from IMDb. In the IMDb dataset, all the data including posters and still frames of each movie can be crawled from the IMDb website. Table 1 shows the statistics of datasets. From Table 1, we can see that the MovieLens dataset has 2113 users, 7258 movies, 8961 posters, and 119162 still frames. The IMDb dataset has 2030 users, 5254 movies, 6492 posters, and 87493 still frames. We split each dataset by randomly assigning 70% to training set and the rest 30% to test set.

B. EVALUATION METRICS

Since our proposed model UVMF-AES is based on probabilistic matrix factorization, the model performance is measured by computing Root Mean Square Error (abbreviated as *RMSE*) on the test set. We adopt *RMSE* as an evaluation criterion, and the formula is as follows,

$$RMSE = \sqrt{\sum_{(u,v)\in Z} \frac{(r_{uv} - \hat{r}_{uv})^2}{|Z|}} \tag{15}$$

where $|Z|$ is the number of test ratings, r_{uv} means the observed score user u rated on movie v, and $\hat{r}_{uv}$ represents the predicted score user u rated on movie v from our model UVMF-AES.

After training and optimizing our model, we can get the predicted matrix. Then we rank the predicted values and recommend the top-N movies to each user.

In recommender system, *Precision* and *Recall* are commonly used to evaluate the accuracy of recommendation. Here, we define a movie whose value is lower than 2.5 as a false sample(user dislike it), and the higher one as a true sample(user like it). From Table 2, we use *TP* to denote the number of movies that a user like (True) and we recommend (Positive), *TN* to denote the number of movies that a user like (True) and we don't recommend (Negative), *FP* to denote the number of movies that a user dislike (False) and we recommend (Positive), *FN* to denote the number of movies that a user dislike (False) and we don't recommend (Negative). The formulas of *Precision* and *Recall* are as follows,

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{16}$$

TABLE 1. Data statistics on two real-world datasets.

Dataset	MovieLens	IMDb
nums of users	2113	2030
nums of movies	7258	5254
nums of posters	8961	6492
nums of still frames	119162	87493
rating sparsity	95.3%	94.6%

TABLE 2. Results in classification problem.

	Positive	Negative
True	TP	TN
False	FP	FN

C. COMPARISON AND PERFORMANCE

We compare our proposed model UVMF-AES with five representative methods for movie recommendation.

- PMF: Probabilistic Matrix Factorization [31] explains the latent factor of user and item from the process of probability generation.
- BPR: Bayesian Personalized Ranking [39] is a ranking algorithm based on matrix factorization. It is not a global optimization, but a ranking optimization for user preference.
- CMF: Collective Matrix Factorization [13] alleviates data sparsity and cold-start problem by taking side information into consideration. Here, we consider visual contents and user-item rating.

[1] https://grouplens.org/datasets/movielens/
[2] https://www.imdb.com

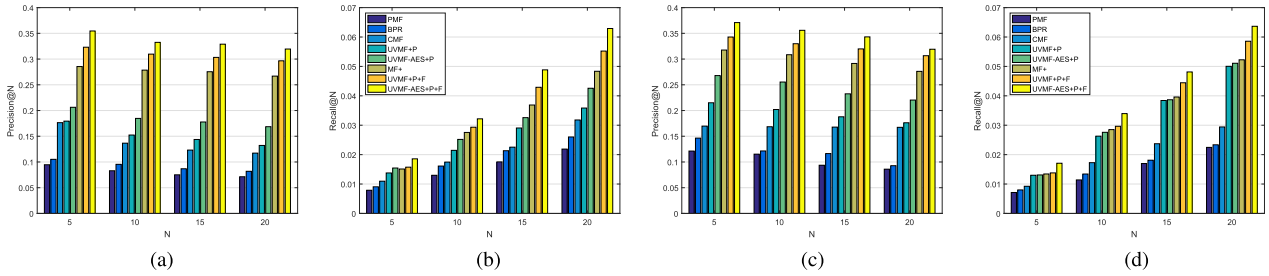

FIGURE 5. The *Precision* and *Recall* of different methods by varying *N*. (a) Precision on MovieLens. (b) Recall on MovieLens. (c) Precision on IMDb. (d) Recall on IMDb.

TABLE 3. Performance comparison of different methods in terms of *RMSE*.

Method	RMSE on MovieLens	RMSE on IMDb
PMF	0.792127	0.772313
BPR	0.785073	0.769193
CMF	0.763221	0.754288
UVMF+P	0.740569	0.735592
UVMF-AES+P	**0.729461**	**0.731458**
MF+	0.726724	0.720632
UVMF+P+F	0.710189	0.706921
UVMF-AES+P+F	**0.691392**	**0.687361**

- MF+: MF+ [4] utilizes image features as a complement to context-aware information and further extends the MF model.
  MF+ model first classifies the image features and then calculates the similarities between these classified features.
- UVMF [5]: UVMF extends probabilistic matrix factorization and considers visual contents for movie recommendation. It builds a bridge between probabilistic matrix factorization and convolutional neural network, and can be trained end-to-end.

The performance of our proposed model UVMF-AES on two real-world datasets evaluated by *RMSE*, *Precision* @N, *Recall* @N is shown in Table 3 and Figure 5. We train other baseline methods on the same training set and evaluate them on the same test set. We set $K = 10$, which is the dimension of user and movie features. The learning rate η is set to 0.001 in our experiments. The number of negative samples J is set to 5. The parameter α is set to 0.0001, which controls the contribution of images. The parameters λ_1 and λ_2 are both set to 1 for preventing over-fitting. And we follow the best parameter settings for other baseline methods.

From the experimental results, we give the observations as follows. First, the visual methods UVMF-AES, UVMF, MF+, and CMF performs better than the non-visual methods PMF and BPR, owing to their capabilities in modeling the visual contents. Second, among the visual contents aware methods, UVMF-AES+P+F, UVMF+P+F, and MF+ all perform better than UVMF-AES+P and UVMF(+P). It proves the effectiveness of the idea of considering still frames of movies. In movie recommendation, still frames of movies do help to alleviate data sparsity. Third, UVMF-AES+P+F and UVMF+P+F improve the performance in comparison with MF+ due to the recommendation mission of visual contents. Here, MF+ also utilizes the visual features of movie posters and still frames. Especially, image features are taken into a neighborhood model and are utilized to calculate the similarities between these images features in MF+ model. Since visual features are mainly designed for classification, they are separated from movie features. Hence, they cannot reflect the properties of movies well. Modeling the visual contents of movie posters and still frames, both UVMF-AES+P+F and UVMF+P+F integrate visual feature extraction and recommendation into a unified framework. In the learning process, they have fine-tuned CNN features. It proves that the unified optimization method of visual features and recommendation performs better than those optimized for image classification.

D. EFFECTIVENESS OF AESTHETIC FEATURES

To analyze the effectiveness of aesthetic features, we conduct experiments on the proposed model UVMF-AES. Figure 5 shows that *Precision* and *Recall* of UVMF-AES+P+F are considerably higher than those of UVMF+P+F. From Table 3, *RMSE* of UVMF+P+F is about 0.018 higher than that of UVMF-AES+P+F. Moreover, *RMSE* of UVMF+P is about 0.011 less than that of UVMF-AES+P on MovieLens dataset. Certainly, UVMF-AES, which considers the aesthetic features of images, has better performance. Whatever considering different combinations of movie posters and still frames, UVMF-AES is superior on the same dataset. Here, UVMF only utilizes CNN features. As a part of visual features, aesthetic features enrich the properties of movies. Further, aesthetic features alleviate data sparsity. CNN features describe the content of images and help to find similar movies by the similarity of images. Aesthetic features build a closer relationship between the images which not only have similar contents but also high-level aesthetic assessment. Enriching the visual contents, aesthetic features improve the accuracy of movie recommendation.

VII. CONCLUSION

In the paper, we exploited visual contents of movie posters and still frames to alleviate the data sparsity issue in movie

recommendation. CNN features and aesthetic features could be extracted from visual contents using deep learning networks. We used VGG16 model to extract CNN features, which were used to analyze what the movie is. And we used OWACNN model to extract aesthetic features, which could tell us whether it was good-looking. Further, we integrated these two features into Probabilistic Matrix Factorization, and proposed a novel model named Aesthetic-aware Unified Visual Contents Matrix Factorization (UVMF-AES). Finally, experimental results showed that UVMF-AES outperformed other baseline methods and demonstrated aesthetic features could improve the accuracy of movie recommendation.

REFERENCES

[1] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 931–940.

[2] H. Ma, T. C. Zhou, M. R. Lyu, and I. King, "Improving recommender systems by incorporating social contextual information," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, p. 9, Apr. 2011.

[3] Z. Lu, Z. Dou, J. Lian, X. Xie, and Q. Yang, "Content-based collaborative filtering for news topic recommendation," in *Proc. AAAI*, 2015, pp. 217–223.

[4] L. Zhao, Z. Lu, S. J. Pan, and Q. Yang, "Matrix factorization+ for movie recommendation," in *Proc. IJCAI*, 2016, pp. 3945–3951.

[5] X. Chen *et al.*, "Exploiting visual contents in posters and still frames for movie recommendation," *IEEE Access*, vol. 6, pp. 68874–68881, 2018.

[6] X. Chen, Y. Zhang, Q. Ai, H. Xu, J. Yan, and Z. Qin, "Personalized key frame recommendation," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 315–324.

[7] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 391–400.

[8] P. Zhao, X. Xu, Y. Liu, V. S. Sheng, K. Zheng, and H. Xiong, "Photo2trip: Exploiting visual contents in geo-tagged photos for personalized tour recommendation," in *Proc. ACM Multimedia Conf.*, 2017, pp. 916–924.

[9] W.-H. Jeong, S.-J. Kim, D.-S. Park, and J. Kwak, "Performance improvement of a movie recommendation system based on personal propensity and secure collaborative filtering," *J. Inf. Process. Syst.*, vol. 9, no. 1, pp. 157–172, 2013.

[10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[11] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, vol. 119, L. D. Raedt and S. Wrobel, Eds. Bonn, Germany, Aug. 2005, pp. 713–719. doi: 10.1145/1102351.1102441.

[12] B. M. Marlin, "Modeling user rating profiles for collaborative filtering," in *Advances in Neural Information Processing Systems*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. Vancouver, BC, Canada: MIT Press, Dec. 2003, pp. 627–634. [Online]. Available: http://papers.nips.cc/paper/2377-modeling-user-rating-profiles-for-collaborative-filtering

[13] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Y. Li, B. Liu, and S. Sarawagi, Eds., Las Vegas, NV, USA, Aug. 2008, pp. 650–658. doi: 10.1145/1401890.1401969.

[14] C. Christakou, S. Vrettos, and A. Stafylopatis, "A hybrid movie recommender system based on neural networks," *Int. J. Artif. Intell. Tools*, vol. 16, no. 5, pp. 771–792, 2007.

[15] J. Salter and N. Antonopoulos, "CinemaScreen recommender agent: Combining collaborative and content-based filtering," *IEEE Intell. Syst.*, vol. 21, no. 1, pp. 35–41, Jan. 2006.

[16] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.

[17] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 627–636.

[18] R. He, C. Lin, J. Wang, and J. McAuley. (2016). "Sherlock: Sparse hierarchical embeddings for visually-aware one-class collaborative filtering." [Online]. Available: https://arxiv.org/abs/1604.05813

[19] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 43–52.

[20] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 507–517.

[21] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retr.*, 2013, pp. 105–112.

[22] X. Li, T.-A. N. Pham, G. Cong, Q. Yuan, X.-L. Li, and S. Krishnaswamy, "Where you instagram?: Associating your instagram photos with points of interest," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1231–1240.

[23] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.

[24] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. World Wide Web Conf. World Wide Web*. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 649–658.

[25] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 497–506.

[26] Y. Q. Zhou, G. Wu, S. Sanner, and P. Manggala. (2018). "Aesthetic features for personalized photo recommendation." [Online]. Available: https://arxiv.org/abs/1809.00060

[27] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[28] P. Lv *et al.* (2018). "USAR: An interactive user-specific aesthetic ranking framework for images." [Online]. Available: https://arxiv.org/abs/1805.01091

[29] Y. Deng, C. C. Loy, and X. Tang, "Aesthetic-driven image enhancement by adversarial learning," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, 2018, pp. 870–878.

[30] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proc. ACM Multimedia Conf.*, 2018, pp. 879–886.

[31] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1257–1264.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015. doi: 10.1007/s11263-015-0816-y.

[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2014, pp. 806–813.

[35] K.-H. Lu, K.-Y. Chang, and C.-S. Chen, "Image aesthetic assessment via deep semantic aggregation," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2016, pp. 232–236.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[39] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 452–461.

XIAOJIE CHEN is currently pursuing the master's degree with the School of Computer Science and Technology, Soochow University, Suzhou. His main research interests include spatial data processing, recommender systems, and data mining.

PENGPENG ZHAO received the Ph.D. degree in computer science from Soochow University, in 2008, where he is currently an Associate Professor and a leading talent in Suzhou Industrial Park. He has published more than 50 papers in prestigious international conferences and journals, including ACM MM, IJCAI, ICDM, CIKM, DASFAA, ICME, and WWWJ. His current research interests include data mining, deep learning, big data analysis, and recommender systems. He has served as an editorial board member of the *International Journal of Intelligent Information Systems*, a reviewer of many important academic journals, such as *ACM Transactions on Knowledge Discovery from Data*, the IEEE TRANSACTIONS ON BIG DATA, WWWJ, and *Neurocomputing*, and a program committee member of international conferences, such as AAAI, IJCAI, CIKM, and PAKDD.

YANCHI LIU received the B.S. degree from the Civil Aviation University of China, and the Ph.D. degree from the University of Science and Technology Beijing. He is currently pursuing the Ph.D. degree from the Management Science and Information Systems Department, Rutgers University, the State University of New Jersey. His research interests include data mining, urban computing, business intelligence, and recommender systems.

LEI ZHAO received the Ph.D. degree in computer science from Soochow University, in 2006, where he is currently a Professor with the School of Computer Science and Technology. His researches focus on graph databases, social media analysis, query outsourcing, and parallel and distributed computing. His recent research is to analyze large graph database in an effective, efficient, and secure way. He has published more than 100 papers, including more than 20 published in well-known journals and conferences, such as ICDE, DASFAA, WISE, and JCST.

JUNHUA FANG received the Ph.D. degree from East China Normal University, in 2017, under the supervision of Prof. A. Zhou. He has been a Lecturer with the Advanced Data Analytics Group, School of Computer Science and Technology, Soochow University, working with Prof. X. Zhou, since 2017. His research interests include distributed stream processing, cloud computing, and spatio-temporal database.

VICTOR S. SHENG received the M.S. degree in computer science from the University of New Brunswick, New Brunswick, in 2003, and the Ph.D. degree in computer science from the Western University, London, ON, Canada, in 2007. After receiving the Ph.D. degree, he was an Associate Research Scientist and NSERC Postdoctoral Fellow in information systems with the Stern Business School, New York University, New York, NY, USA. He is currently an Assistant Professor of computer science with the University of Central Arkansas, Conway, and the Founding Director of Data Analytics Laboratory. His research interests include data mining, machine learning, and related applications.

ZHIMING CUI is currently a Professor with the Institute of Intelligent Information Processing and Application, Soochow University, China. He is also an Outstanding Expert of Jiangsu Province, China. He presided four National Natural Science Foundation of China. He has published several articles in computer vision, data mining, image processing, and pattern recognition. His research interests include deep web, computer vision, image processing, and pattern recognition.