Low-Rank Multi-View Embedding Learning for Micro-Video Popularity Prediction

Peiguang Jing, Yuting Su, Liqiang Nie, *Member, IEEE*, Xu Bai,
Jing Liu, *Member, IEEE*, and Meng Wang, *Member, IEEE*

Abstract—Recently, a prevailing trend of user generated content (UGC) on social media sites is the emerging micro-videos. Micro-videos afford many potential opportunities ranging from network content caching to online advertising, yet there are still little efforts dedicated to research on micro-video understanding. In this paper, we focus on popularity prediction of micro-videos by presenting a novel low-rank multi-view embedding learning framework. We name it as transductive low-rank multi-view regression (TLRMVR), and it is capable of boosting the performance of micro-video popularity prediction by jointly considering the intrinsic representations of the source and target samples. In particular, TLRMVR integrates low-rank multi-view embedding and regression analysis into a unified framework such that the lowest-rank representation shared by all views not only captures the global structure of all views, but also indicates the regression requirements. The framework is formulated as a regression model and it seeks a set of view-specific projection matrices with low-rank constraints to map multi-view features into a common subspace. In addition, a multi-graph regularization term is constructed to improve the generalization capability and further prevents the overfitting problem. Extensive experiments conducted on a publicly available dataset demonstrate that our proposed method achieve promising results as compared with state-of-the-art baselines.

Index Terms—Low-rank learning, multi-view fusion, subspace learning, popularity prediction, micro-video analysis

1 Introduction

Recently, micro-videos, or online short videos, have emerged as a new trend in user-generated content, and such videos have been widely spreading across various social platforms. Generally, micro-videos are created by individuals and contain some unique characteristics, including egocentric and self-facing views. The length of micro-video is always limited. Vine[1] and Snapchat,[2] for example, shorten the length to less than 6 and 10 seconds, respectively. Despite their shortness, micro-videos generally outline a relatively simple but complete story to audiences. Within a limited time interval, producers also attempt to condense and maximize what they want to say, thereby to create more attractive stories. Compared with traditional videos like the ones in Youtube,[3] micro-videos are produced to satisfy a fast-paced modern society, which makes micro-videos appear to be more social-oriented.

Several pioneer efforts have been dedicated to micro-video analysis studies [1], [2], [3], [4]. In particular, Redi et al. [1] studied the creativity of micro-videos by analyzing the audio-visual features that make a video creative. Zhang et al. [2] proposed a novel multi-task multi-modal algorithm to address venue category estimation of micro-videos. Chen et al. [3] presented a transductive multi-modal learning method to predict the popularity of micro-videos. Nguyen et al. [4] constructed a novel micro-video dataset as well as introduced viewpoint-specific and temporally evolving models for micro-video understanding. Motivated by the trend of micro-videos in both academia and industry, in this paper, we work towards solving the problem of popularity prediction of micro-videos posted on social networks.

Popularity prediction of micro-videos is able to benefit many potential applications, ranging from network content monitoring and dynamic bandwidth management to micro-video advertising services. Since micro-videos are produced with the aim of rapid spreading and sharing among users, these videos bring more intrinsic relations with social networks that differ from traditional long videos, therefore it makes predicting the popularity of micro-videos a non-trivial task due to the following facts: 1) *Heterogeneous*: A micro-video can be comprehensively represented by exploiting a combination of visual, acoustic, textual, and social features, whereby one of the critical issues arising in real-world applications is the heterogeneous gap among features extracted from distinct views. Under this scenario, traditional methods that fuse these features using simple concatenation or feature selection approaches may not work well in capturing the semantic understanding of features and may hence lead to information redundancy at the learning stage. It is hence highly desired to comprehensively

[1] https://vine.co.
[2] https://www.snapchat.com.
[3] www.youtube.com.

- *P. Jing, Y. Su, X. Bai, and J. Liu are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. E-mail: {pgjing, ytsu, bayminbx, jliu_tju}@tju.edu.cn.*
- *L. Nie is with the School of Computer Science and Technology, Shandong University, Jinan, Shandong 250000, China. E-mail: nieliqiang@gmail.com.*
- *M. Wang is with the School of Computer and Information Science, Hefei University of Technology, Hefei 230009, China. E-mail: eric.mengwang@gmail.com.*

Manuscript received 1 June 2017; revised 26 Oct. 2017; accepted 11 Dec. 2017. Date of publication 21 Dec. 2017; date of current version 5 July 2018.
(Corresponding author: Yuting Su.)
Recommended for acceptance by S. Yan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2785784

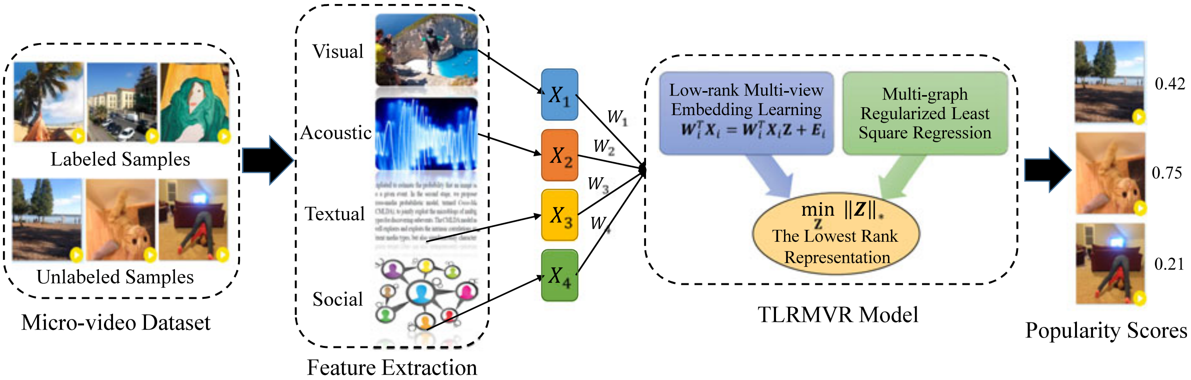


Fig. 1. An illustration of our proposed scheme, consisting of two main components: low-rank multi-view embedding learning and multi-graph regularized least square regression.

consider the intrinsic semantic structure for all heterogeneous features in a unified framework. 2) *Interconnected*: Heterogeneous features extracted from different views show different aspects of micro-videos, which are complementary to each other. In this case, it will be beneficial to develop an effective approach to finding the interconnected patterns shared by all views. However, due to the restrictions brought by micro-video producers and platforms, the additional information associated with a micro-video, textual description, for example, suffers from the diverse or unstructured nature, causing the features extracted from certain views unavailable in many situations. For example, according to the statistics [3], more than 11 percent of micro-videos do not provide textual descriptions. In contrast, micro-video content itself ensures a steady information source to enable popularity prediction. Thus, to compensate for this limitation, micro-video content features are considered to be an indispensable component for a more descriptive and predictive analysis on the one hand, and it is necessary to exploit the complementarity between different views to learn the latent interconnected patterns to address the incomplete problem on the other hand. 3) *Noisy*: Originating from certain external factors in reality, various types of noises make the real underlying data structure hidden in the observed data. For example, micro-videos are often captured by users with hand-held mobile devices which easily result in poor video quality, such as low-resolution, wobbly frames, constrained lighting conditions, and background noise. Besides, textual descriptions related to micro-videos may be noisy and uncorrelated. The aforementioned challenges drive us to build a robust model to explore the intrinsic structure property embedded in data by inferring meaningful features and alleviating the impact of noisy ones.

To address the challenges presented above, in this paper, we first employ four types of heterogeneous features to comprehensively characterize various aspects of micro-videos including the visual, acoustic, social, and textual modalities. As the popularity scores of micro-videos are continuous, we then formulate the task of micro-video popularity prediction as a regression problem and propose a novel low-rank multi-view learning framework named transductive low-rank multi-view regression (TLRMVR). TLRMVR is able to learn a latent common subspace to fuse all the multi-view features such that the lowest-rank representations of the source and target are obtained for micro-video popularity prediction. Fig. 1 illustrates the scheme of TLRMVR, comprising of two main components: low-rank multi-view embedding learning and multi-graph regularized least squares regression. The goal of the former is to learn a set of view-specific transformation matrices by maximizing the total correlations between any two views with low-rank constraints. Due to the strength of the low-rank constraint in addressing incomplete and noisy information, low-rank representation has been successfully applied to a wide range of applications, such as subspace segmentation [5], [6], [7] and visual classification [8], [9], [10]. We are inspired to integrate the advantages of both low-rank representation and multi-view learning to enhance the robustness of feature learning. As to the second component, it is to build connections between latent low-rank representations and popularity scores. In this regard, a multi-graph regularization is constructed to improve the generalization performance and prevent overfitting. By unifying the low-rank representation and regression analysis, the lowest-rank representations shared by all views not only capture the global structure of all heterogeneous features but also indicate the regression requirements. Because the formulated objective function is non-smooth and hard to solve, we design an effective algorithm based on the augmented Lagrange multiplier (ALM) to optimize it and ensure a fast convergence.

We summarize the main contributions of our work as follows:

- We present a transductive low-rank multi-view regression framework for micro-video popularity prediction. Under the proposed framework, micro-videos are encoded from diverse perspectives with the aim of representing micro-videos as comprehensively as possible.
- TLRMVR integrates the advantages of low-rank representation and multi-view learning to address the heterogeneous, interconncected, and noisy data problems. To the best of our knowledge, this work is the first attempt to jointly integrate low-rank representation and multi-view learning into the task of micro-video popularity prediction.
- We develop an alternating iterative algorithm by applying the augmented Lagrangian multiplier method to optimize our model. The experiments conducted on a publicly available micro-video dataset confirm the convergence and effectiveness of our proposed scheme.

The remainder of this paper is organized as follows. In Section 2, we first briefly review pioneering efforts related to low-rank subspace learning and popularity prediction. Then, our proposed model is presented in Section 3. Experimental evaluation and analysis of our method is reported in Section 4, followed by conclusion and future work in Section 5.

2 Related Work

2.1 Low-Rank Subspace Learning

In recent years, low-rank representation [5], [11], [12], [13] has been considered as a promising technique for exploring the latent low-dimensional representation embedded in the original space. Low-rank subspace learning has been applied to a wide range of machine learning tasks, including matrix recovery [14], image classification [9], [10], subspace segmentation [6], and missing modality recognition [15].

Robust principal component analysis (RPCA) [16] is a popular low-rank matrix recovery method for high-dimensional data processing. This method aims to decompose a data matrix into a low-rank matrix and a sparse error matrix. To promote the discriminative ability of the original RPCA and improve the robust representation of corrupted data, Chen et al. [17] presented a novel low-rank matrix approximation method with a structural incoherence constraint, which decomposes the raw data into a set of representative bases with associated sparse error matrices. Based on the principle of self-representation, Liu et al. [5] proposed the low-rank representation (LRR) method to search for the lowest-rank representation among all the candidates. To overcome the incompetence of LRR in handling unobserved, insufficient and extremely noisy data, Liu and Yan [6] further developed an advanced version of LRR, called latent low-rank representation (LatLRR), for subspace segmentation. Zhang et al. [18] proposed a structured low-rank representation method for image classification, which constructs a semantic-structured and constructive dictionary by incorporating class label information into the training stage. Zhou et al. [19] provided a novel supervised and low-rank-based discriminative feature learning method that integrates LatLRR with ridge regression to minimize the classification error directly.

To handle data that are generated from multiple views in many real-world applications, some multi-view low-rank subspace learning methods have been developed to search for a latent low-dimensional common subspace such that it can capture the commonality among all the views. For example, Xia et al. [20] proposed to construct a transition probability matrix from each view and then recover a shared low-rank transition probability matrix via low-rank and sparse decomposition. Liu et al. [21] presented a novel low-rank multi-view matrix completion (lrMMC) method for multi-label image classification, where a set of basic matrices are learned by minimizing the reconstruction errors and the rank of the latent common representation. In the case that the view information of the testing data is unknown, Ding and Fu [22] proposed a novel low-rank common subspace (LRCS) algorithm in a weakly supervised setting, where only the view information is employed in the training phase. In [23], a dual low-rank decomposition model was developed to learn a low-dimensional view-invariant subspace. To guide the decomposition process, two supervised graph regularizers were considered to separate the class structure and view structure. Li et al. [24] proposed a novel approach, named low-rank discriminant embedding (LRDE), by considering the correlations between views and the geometric structures contained within each view simultaneously. These multi-view low-rank learning approaches have been proven to be effective when different feature views are complementary to each other.

Although low-rank representation enables an effective learning mechanism in exploring the low-rank structure in noisy datasets [25], only a limited amount of low-rank models have been developed to address the popularity prediction in social networks. The prediction of video popularity can be considered as a standard regression problem. To the best of our knowledge, one of the most related work to our approach is introduced in [26], in which a multi-view low-rank regression model is presented by imposing low-rank constraints on the multi-view regression model. However, in that work, the structure and relations among different views were ignored. To overcome this drawback, we propose to learn a set of view-specific projections by maximizing the total correlations among views to map multi-view features into a common space. Another difference is that the lowest-rank representation is adaptively obtained by a low-rank constraint, which is approximated by the trace norm rather than being specified in advance.

2.2 Popularity Prediction

Significant efforts were devoted to exploring the popularity prediction of items such as text [27], [28], images [29], [30], [31], and videos [32], [33], [34], [35] due to their potential value in business [36], [37].

For the task of predicting the popularity of text, most methods tend to explore the textual content itself and the correlation between popularity and the social context. For example, Ma et al. [28] proposed to predict the popularity of new hashtags on Twitter by extracting 7 content features from both hashtags and tweets and 11 contextual features from the social graphs formed by users.

As to the image popularity, content-based image features, context features, and social context features are generally exploited to predict image popularity. For example, Khosla et al. [38] explored the relative significance of individual features involving multiple visual features, such as color, gradient, texture, and the presence of objects, as well as various social context features, such as the number of normalized views or contacts. Totti et al. [39] presented a complementary analysis on how the aesthetic properties, such as brightness, contrast and sharpness, and semantics contribute to image popularity. Gelli et al. [31] proposed to combine user features and three context features together with image sentiment features to better predict the popularity of social images.

When it comes to video popularity prediction, analogous to images, videos also integrate different information channels, like visual, acoustic, social, and textual modalities. The majority of studies focus on investigating the factors that determine the popularity of videos [32], [33], [35], [40]. For example, Cha et al. [32] conducted a large-scale data-driven analysis to uncover the latent correlations between video popularity and user-generated content. Li et al. [33]

proposed to use both video attractiveness and social context as inputs to predict video views on online social networks. Trzcinski et al. [35] employed temporal and visual cues to predict the popularity of online videos. The tasks above share the same thing-they do not describe each item based on its content only; instead, they mine multiple views of context information related to the item and social cues from the users to improve the prediction performance. In addition, some works [33], [3], [41] have explored the improvement of video popularity prediction by fusing information introduced by different patterns. To overcome the ineffectiveness of traditional models, such as autoregressive integrated moving average (ARIMA), multiple linear regression (MLR), and k-nearest neighbors (kNN), when predicting the popularity of online videos, Li et al. [33] introduced a novel propagation-based popularity prediction method by considering both video intrinsic attractiveness and the underlying propagation structure. Roy et al. [42] used transfer learning to model the social prominence of videos, in which an intermediate topic space is constructed to connect the social and video domains. Ding et al. [41] developed a dual sentimental Hawkes process (DSHP) for video popularity prediction, which not only takes sentiments in video popularity propagation into account but also reveals more underlying factors that determine the popularity of a video. Moreover, Chen et al. [3] developed a transductive multi-modal learning model (TMALL) to predict the popularity of micro-videos, in which different modal features can be described in a latent common space to solve the noise and insufficiency issues.

However, the aforementioned works do not consider the combined impact of heterogeneous, interconnected, and noisy data. In contrast, our proposed scheme not only pursues a solid fusion of heterogeneous multi-view features based on the complementary characteristics but also concentrates on exploiting the advantages of the low-rank representation to learn robust features within the incomplete and noisy data.

3 Proposed Methodology

3.1 Notations and Preliminaries

In this section, we begin with a brief summary of the involved basic notations. Except for some specific cases, we represent a column vector with lowercase bold letters, e.g., **m**, and a matrix with uppercase bold letters, e.g., **M**. Moreover, given a matrix $\mathbf{M} \in R^{N\times D}$, its ith row and the ith column of matrix **M** are denoted by $\mathbf{m}^i$ and $\mathbf{m}_i$, respectively. The $\mathcal{L}_{p,q}$-norm of matrix **M** is defined as

$$\|\mathbf{M}\|_{p,q}=\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{D}|M_{ij}|^p\right)^{q/p}\right]^{1/q}, \tag{1}$$

where M_{ij} is the (i,j)th element of matrix **M**. By assigning different values to p and q, there are several regularization terms, which are stated as follows.

The $\mathcal{L}_1$-norm is defined when $p=q=1$,

$$\|\mathbf{M}\|_1=\sum_{i=1}^{N}\|\mathbf{m}^i\|_1. \tag{2}$$

The Frobenius norm $\mathcal{L}_F$ is defined when $p=q=2$,

$$\|\mathbf{M}\|_F=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{D}\|M_{ij}\|^2}. \tag{3}$$

The trace norm of matrix **M** is defined as

$$\|\mathbf{M}\|_*=\sum\nolimits_i\delta_i(\mathbf{M}), \tag{4}$$

where $\sum_i\delta_i(\mathbf{M})$ is the sum of singular values of matrix **M**.

3.2 Problem Formulation

Considering that we collect N samples labeled with popularity scores and M unlabeled samples, we extract K types of feature sets for this collection; hence, we obtain the feature matrix $\mathbf{X}=[\mathbf{X}_1;\mathbf{X}_2;\cdots;\mathbf{X}_K]$, where $\mathbf{X}_i \in R^{D_i\times(N+M)}$ encodes the ith feature type and D_i is the dimensionality corresponding to the ith feature type. Without loss of generality, we assume that columns of $\mathbf{X}_i$ are mean centered. Meanwhile, we denote $\mathbf{y}=[y_1,\ldots,y_N,0,\ldots,0]^T \in R^{N+M}$ as the popularity score vector for all samples, where y_i is the popularity score of the ith sample.

3.2.1 Low-Rank Multi-View Embedding Learning

Traditional canonical correlation analysis (CCA) [43] aims to find a common subspace in which two views of variables are fused with the maximum correlation assumption. Inspired by the success of CCA, multi-view canonical correlation analysis (MCCA) [44] was developed as a generalized CCA for multi-view scenarios. Specifically, to fully exploit the complementary properties of different views to eliminate the heterogeneity among them, MCCA attempts to find multiple basic transformation matrices $\{\mathbf{W}_i|\mathbf{W}_i \in R^{D_i\times D}\}_{i=1}^K$ with $1\leq D\leq \min(D_1,D_2,\ldots,D_K)$ to respectively project the samples in the K views to an intrinsic low-dimensional space such that the total correlation across all view pairs is maximized while partly discarding the redundancy. Formally, it can be formulated as,

$$\begin{aligned}&\max_{\mathbf{W}_1,\mathbf{W}_2,\ldots,\mathbf{W}_M}\sum_{i=1}^{K}\sum_{j=1}^{K}tr(\mathbf{W}_i^T\mathbf{S}_{ij}\mathbf{W}_j)\\ &s.t.\ \mathbf{W}_i^T\mathbf{S}_{ii}\mathbf{W}_i=\mathbf{I}_{D\times D},\ \ i=1,\ldots,K,\end{aligned} \tag{5}$$

where $\mathbf{I}_{D\times D}$ is a $D\times D$ identity matrix, $\mathbf{S}_{ij}\in R^{D_i\times D_j}$ is defined as covariance matrices of $\mathbf{X}_i$ and $\mathbf{X}_j$. In a compact form, the total correlation in the common space in Eq. (5) can be reformulated as follows:

$$\max_{\mathbf{W}_1,\ldots,\mathbf{W}_K}\ tr(\hat{\mathbf{W}}^T\mathbf{S}\hat{\mathbf{W}})\quad s.t.\ \hat{\mathbf{W}}^T\hat{\mathbf{S}}\hat{\mathbf{W}}=\mathbf{I}_{D\times D}, \tag{6}$$

where $\hat{\mathbf{W}}=[W_1;\mathbf{W}_2;\cdots;\mathbf{W}_K]\in R^{(D_1+D_2+\cdots+D_K)\times D}$; $\hat{\mathbf{S}}=\text{diag}\{\mathbf{S}_{11},\mathbf{S}_{22},\ldots,\mathbf{S}_{KK}\}\in R^{(D_1+\cdots+D_K)\times(D_1+\cdots+D_K)}$ is a block-diagonal matrix; and **S** is the block matrix of the same size as $\hat{\mathbf{S}}$

$$\mathbf{S}=\begin{pmatrix}\mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1K}\\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2K}\\ \cdots & \cdots & \cdots & \cdots\\ \mathbf{S}_{K1} & \mathbf{S}_{K2} & \cdots & \mathbf{S}_{KK}\end{pmatrix}.$$

The low-rank constraint is helpful for finding a more robust subspace structure and meanwhile removing the noise information from the data. Considering the aforementioned advantages, LRR assumes that the feature matrix can be decomposed into a salient part and a sparse error part. Following the scheme in LRR, we factorize the multiple view-specific transformed matrices $\mathbf{W}_1^T\mathbf{X}_1, \mathbf{W}_2^T\mathbf{X}_2, \ldots, \mathbf{W}_K^T\mathbf{X}_K$ into salient parts with a latent common low-rank structure $\mathbf{Z}$ shared by all views and their unique error matrices $\mathbf{E}_1$, $\mathbf{E}_2, \ldots, \mathbf{E}_K$. To better separate the salient part and the error part, we need to solve the following optimization problem

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E}_i} \ \mathrm{rank}(\mathbf{Z}) + \lambda \sum_{i=1}^{K} \|\mathbf{E}_i\|_1 \\ &s.t.\ \mathbf{W}_i^T\mathbf{X}_i = \mathbf{W}_i^T\mathbf{X}_i\mathbf{Z} + \mathbf{E}_i, \ \ i = 1, \ldots, K, \end{aligned} \tag{7}$$

where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N+M}] \in R^{(N+M)\times(N+M)}$ is a common low-rank representation of all samples for the view-variance structure; $\mathbf{E}_i \in R^{D\times(N+M)}$ is the unique sparse error part constrained by the $\mathcal{L}_1$-norm to handle random corruption; and $\lambda > 0$ is a balanced parameter.

There are two different understandings of the low-rank matrix $\mathbf{Z}$: one is that the matrix $\mathbf{Z}$ can be considered as an affinity matrix whose elements Z_{ij} reflect the similarity between the ith and the jth samples, and the other is that the feature vectors corresponding to samples of the columns of the feature matrix $\mathbf{Z}$, which plays a dominant role in representing structures learned from multiple views.

Since Eq. (7) is difficult to be optimized due to the nonconvex $\mathrm{rank}(\cdot)$, the nuclear norm $\|\mathbf{Z}\|_*$ is adopted to approximate the rank of matrix $\mathbf{Z}$. Thus, the result of Eq. (7) can be derived in a compact form as

$$\min_{\mathbf{Z},\,\mathbf{E}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_1 \ s.t.\ \hat{\mathbf{W}}^T\mathbf{X} = \hat{\mathbf{W}}^T\mathbf{X}\mathbf{Z} + \mathbf{E}, \tag{8}$$

where $\mathbf{E} = \left[E_1^T, \mathbf{E}_2^T, \ldots, \mathbf{E}_K^T\right]^T \in R^{KD\times(N+M)}$.

Therefore, the objective function of the low-rank multi-view embedding learning can be obtained by combining the objective functions in Eqs. (6) and (8),

$$\begin{aligned} &\min_{\mathbf{Z},\,\mathbf{E},\hat{\mathbf{W}}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_1 - \delta tr(\hat{\mathbf{W}}^T\mathbf{S}\hat{\mathbf{W}}) \\ &s.t.\ \hat{\mathbf{W}}^T\mathbf{X} = \hat{\mathbf{W}}^T\mathbf{X}\mathbf{Z} + \mathbf{E}, \ \ \hat{\mathbf{W}}^T\hat{\mathbf{S}}\hat{\mathbf{W}} = \mathbf{I}, \end{aligned} \tag{9}$$

where δ is a balanced parameter.

From another perspective, the core idea of the low-rank multi-view embedding learning is to utilize the relationship between features and samples. At the sample level, with the assumption that samples from different views are correlated with each other, a set of view-specific transformation matrices are obtained to project the multi-view features into a common low-dimensional subspace. At the feature level, with the greater robustness of low-rank constraint to data noise, the lowest-rank representation of each sample is obtained by revealing the underlying low-rank subspace structure spanned by the transformed samples.

3.2.2 Multi-Graph Regularized Least Squares Regression

As the popularity scores of micro-videos are continuous, from a narrower sense, regression analysis refers specifically to the estimation of continuous variables, which is opposite to the discrete variables used in classification. Regression analysis is widely used for predicting and forecasting tasks [31], [35], [45], [46], [47]. For example, Gelli et al. [31] modeled image popularity prediction as a $\mathcal{L}_2$ regularized $\mathcal{L}_2$ loss support vector regression (SVR) problem. Szabo et al. [46] presented a linear regression method with the maximum likelihood to predict online content popularity. Trzcinski et al. [35] introduced a SVR method with Gaussian radial basis functions (RBF) to predict the popularity of online videos. Similarity, Peng et al. [47] addressed the issue of image memorability prediction by proposing a multi-view adaptive regression model.

In this section, we use the same approach and regard the popularity prediction of micro-videos as a regression problem. For simplicity and efficiency in solving this problem, we adopt the commonly used ordinary least squares (OLS) approach, which considers a linear dependence $\mathbf{w}$ between the input feature matrix $\mathbf{Z}$ and the output popularity score $\mathbf{y}$. After adding a ridge regularization to the least squares loss part $\|\mathbf{y} - \mathbf{Z}^T\mathbf{w}\|_2^2$, we obtain a typical least squares problem with ridge regression,

$$\min_{\mathbf{w},\mathbf{Z}} \frac{1}{2}\left\|\mathbf{y} - \mathbf{Z}^T\mathbf{w}\right\|_2^2 + \alpha\|\mathbf{w}\|_2^2, \tag{10}$$

where $\mathbf{w} \in R^{N+M}$ is the regression coefficient and α is a parameter to balance the tradeoff between the empirical loss and the regularization penalty.

Furthermore, to provide some leeway for test samples to avoid the predicted results being regressed to zero values, we rewrite the objective function as

$$\min_{\mathbf{w},\mathbf{Z}} \frac{1}{2}\left\|\mathbf{y} - (\mathbf{Z}\mathbf{M})^T\mathbf{w}\right\|_2^2 + \alpha\|\mathbf{w}\|_2^2, \tag{11}$$

where $\mathbf{M}$ is a block diagonal matrix used to select labeled samples from all samples, which is defined by $\mathbf{M} = \begin{bmatrix} \mathbf{I}_{N\times N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in R^{(N+M)\times(N+M)}$.

Since the low-rank embedding is learned to characterize the popularity of micro-videos, to better guide the low-rank multi-view embedding learning while avoiding the overfitting problem of test samples, we also consider the fact that the subspaces spanned by the original features and the predicted results should contain similar local geometric structures. Rather than employing a simple concatenation of features to characterize the geometric structure with a graph Laplacian, we compute a unified graph Laplacian to fuse the structures embedded in different views. Then, the geometrical structure consistency between multi-view features and the smoothness of a vectorial prediction function $f : \mathbf{z} \to R$ on graphs are preserved by minimizing the following regularizer

$$\begin{aligned} \Omega(f) &= \sum_{k=1}^{K} \sum_{i,j=1}^{N+M} \left\|\mathbf{w}^T\mathbf{z}_i - \mathbf{w}^T\mathbf{z}_j\right\|_2^2 S_{ij}^k \\ &= \sum_{k=1}^{K} \mathbf{w}^T\mathbf{Z}(\mathbf{D}^k - \mathbf{S}^k)\mathbf{Z}^T\mathbf{w} \\ &= \sum_{k=1}^{K} \mathbf{w}^T\mathbf{Z}\mathbf{L}^k\mathbf{Z}^T\mathbf{w} \\ &= \mathbf{w}^T\mathbf{Z}\mathbf{L}\mathbf{Z}^T\mathbf{w}, \end{aligned} \tag{12}$$

where $\mathbf{L} = \sum_{k=1}^{K} \mathbf{L}^k$ is a unified Laplacian matrix, $\mathbf{L}^k = \mathbf{D}^k - \mathbf{S}^k$ is the graph Laplacian matrix for the kth view, $\mathbf{S}^k$ is the weight matrix computed by the Gaussian similarity function, and $\mathbf{D}^k$ is the diagonal degree matrix with $D_{ii}^k = \sum_j S_{ij}^k$. Here, $\mathbf{S}^k$ is computed as follows:

$$S_{ij}^k = \begin{cases} \exp\left(-\frac{\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|_2^2}{2\sigma^2}\right) & \text{if } x_i \in N_{\tilde{k}}(x_j) \text{ or } \\ & \quad x_j \in N_{\tilde{k}}(x_i) \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ are the ith and jth samples in the kth feature space, respectively; $x_i \in N_{\tilde{k}}(x_j)$ means that x_i is the $\tilde{k}$ nearest neighbor of data x_j; and σ is the radius parameter, which is simply set as the median of the Euclidean distances over all micro-video pairs.

Therefore, by combining Eqs. (11) and (12), the objective function based on low-rank representation is formulated as follows:

$$\min_{\mathbf{w},\mathbf{Z}} \frac{1}{2}\left\|\mathbf{y} - (\mathbf{ZM})^T\mathbf{w}\right\|_2^2 + \phi\mathbf{w}^T\mathbf{ZLZ}^T\mathbf{w} + \alpha\|\mathbf{w}\|_2^2, \tag{14}$$

where ϕ is a balanced parameter.

To better guide the low-rank subspace learning in our previous model, we develop a quadratic term $\mathcal{G}(\mathbf{w}, \mathbf{Z}, \mathbf{W})$ by combining supervised information (i.e., regression information and view information) and multi-graph regularizer. Based on the above formulations, the quadratic term $\mathcal{G}(\mathbf{w}, \mathbf{Z}, \mathbf{W})$ on all samples is formulated as follows:

$$\begin{aligned} &\mathcal{G}(\mathbf{w}, \mathbf{Z}, \hat{\mathbf{W}}) \\ &= \frac{1}{2}\left\|\mathbf{y} - (\mathbf{ZM})^T\mathbf{w}\right\|_2^2 + \phi\mathbf{w}^T\mathbf{ZLZ}^T\mathbf{w} - \delta tr(\hat{\mathbf{W}}^T\mathbf{S}\hat{\mathbf{W}}). \end{aligned}$$

By combining the objective functions in Eqs. (6) and (14) with Eq. (8), we develop our TLRMVR algorithm as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{Z},\mathbf{E},\hat{\mathbf{W}}} \ & \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_1 + \alpha\|\mathbf{w}\|_2^2 + \beta\mathcal{G}(\mathbf{w}, \mathbf{Z}, \hat{\mathbf{W}}) \\ s.t. \ & \hat{\mathbf{W}}^T\mathbf{X} = \hat{\mathbf{W}}^T\mathbf{XZ} + \mathbf{E}, \hat{\mathbf{W}}\hat{\mathbf{S}}\hat{\mathbf{W}}^T = \mathbf{I}. \end{aligned} \tag{15}$$

Here, we initialize $\hat{\mathbf{W}}$ by the following trace ratio equation

$$\{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K\} = \arg\max_{\mathbf{W}_{1,},\ldots,\mathbf{W}_K} \frac{tr(\hat{\mathbf{W}}^T\mathbf{S}\hat{\mathbf{W}})}{tr(\hat{\mathbf{W}}^T\hat{\mathbf{S}}\hat{\mathbf{W}})}. \tag{16}$$

When the low-rank representation from multi-view embedding learning and regression analysis are both performed, the lowest-rank representation shared by all views not only captures the global structure of all modalities but also indicates the regression requirements.

3.3 Optimization

The objective function in Eq. (15) can be solved by applying the alternating direction method of multipliers (ADMM), which divides a complex problem into subproblems, where each of them is easier to handle with an iterative process. We first introduce two Lagrange multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$ to obtain the so-called augmented Lagrangian function.

Then, we merge the last five terms into a single one: $H(\mathbf{w}, \mathbf{Z}, \mathbf{E}, \hat{\mathbf{W}}, \mathbf{Y}_1, \mathbf{Y}_2, \mu) = \beta\mathcal{G}(\mathbf{w}, \mathbf{Z}, \hat{\mathbf{W}}) + \frac{\mu}{2}\|\hat{\mathbf{W}}^T\mathbf{X} - \hat{\mathbf{W}}^T\mathbf{XZ} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu}\|_F^2 + \frac{\mu}{2}\|\hat{\mathbf{W}}^T\hat{\mathbf{S}}\hat{\mathbf{W}} - \mathbf{I} + \frac{\mathbf{Y}_2}{\mu}\|_F^2$; thus, the augmented Lagrangian function of Eq. (15) can be reformulated as follows:

$$\begin{aligned} & L(\mathbf{w}, \mathbf{Z}, \mathbf{E}, \hat{\mathbf{W}}, \mathbf{Y}_1, \mathbf{Y}_2, \mu) \\ & = \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_1 + \alpha\|\mathbf{w}\|_2^2 - \frac{1}{2\mu}\left(\|\mathbf{Y}_1\|_F^2 + \|\mathbf{Y}_2\|_F^2\right) \\ & \quad + H(\mathbf{w}, \mathbf{Z}, \mathbf{E}, \hat{\mathbf{W}}, \mathbf{Y}_1, \mathbf{Y}_2, \mu). \end{aligned} \tag{17}$$

To better interpret the process, we introduce a variable t and define $\mathbf{Z}_t, \mathbf{E}_t, \mathbf{W}_t, \mathbf{w}_t, \mathbf{Y}_{1,t}, \mathbf{Y}_{2,t}$, and μ_t as the variables updated in the tth iteration. Under the ADMM framework, the problem L with respect to each variable in the $t+1$ iteration is optimized as the following scheme:

For $\mathbf{Z}$: We can update $\mathbf{Z}$ by dropping the terms independent of $\mathbf{Z}$ as the following scheme:

$$\begin{aligned} \mathbf{Z}_{t+1} &= \arg\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + H(\mathbf{Z}, \mathbf{E}, \hat{\mathbf{W}}, \mathbf{Y}_1, \mathbf{Y}_2, \mu) \\ &= \arg\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{\tau_t\mu_t}{2}\|\mathbf{Z} - \mathbf{Z}_t\|_F^2 + \langle \triangledown_{\mathbf{Z}}H, \mathbf{Z} - \mathbf{Z}_t\rangle \\ &= \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Z} - \mathbf{Z}_t + \triangledown_{\mathbf{Z}}H\|_F^2 + \frac{1}{\tau_t\mu_t}\|\mathbf{Z}\|_*, \end{aligned} \tag{18}$$

where $\triangledown_{\mathbf{Z}}H = \beta\mathbf{w}_t(\mathbf{w}_t^T \quad \mathbf{Z}_t\mathbf{M} - \mathbf{y}^T)\mathbf{M}^T + 2\phi\beta\mathbf{w}_t\mathbf{w}_t^T\mathbf{Z}_t\mathbf{L} - \mu_t\mathbf{X}^T\hat{\mathbf{W}}_t(\hat{\mathbf{W}}_t^T\mathbf{X} - \hat{\mathbf{W}}_t^T\mathbf{XZ}_t - \mathbf{E}_t + \frac{1}{\mu_t}\mathbf{Y}_{1,t})$ is the partial derivative $H(\mathbf{Z}, \mathbf{E}, \mathbf{W}, \mathbf{Y}_1, \mathbf{Y}_2, \mu)$ with respect to $\mathbf{Z}$ and $\tau_t = \|\hat{\mathbf{W}}_t^T\mathbf{X}\|_F^2$.

The problem in Eq. (18) is a standard nuclear norm minimization problem, which can be approximately solved by the singular value thresholding (SVT) algorithm [48]. Specifically, suppose that the singular vector decomposition of $\mathbf{Z}_t - \triangledown_{\mathbf{Z}}H$ of rank r is

$$\mathbf{Z}_t - \triangledown_{\mathbf{Z}}H = \mathbf{P}\Sigma\mathbf{Q}^T, \quad \Sigma = \text{diag}(\{\delta_i\}_{i=1}^r), \tag{19}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are left-singular and right-singular matrices with orthogonal columns and Σ is a rectangular diagonal matrix with non-negative real numbers δ_i on the diagonal. Then, the optimal solution $\mathbf{Z}$ is $\mathbf{Z}_{t+1} = \mathcal{D}_{1/\tau_t\mu_t}(\mathbf{Z}_t - \triangledown_{\mathbf{Z}}H)$. For each $1/\tau_t\mu_t \geq 0$, the soft-thresholding operator $\mathcal{D}_{1/\tau_t\mu_t}$ $(\mathbf{Z}_t - \triangledown_{\mathbf{Z}}H)$ is defined as [48]

$$\begin{aligned} \mathcal{D}_{1/\tau_t\mu_t}(\mathbf{Z}_t - \triangledown_{\mathbf{Z}}H) &= \mathbf{P}\Sigma_{\mathbf{1/\tau_t\mu_t+}}\mathbf{Q}^T \\ \Sigma_{\mathbf{1/\tau_t\mu_t+}} &= \text{diag}(\{(\delta_i - 1/\tau_t\mu_t)_+\}), \end{aligned} \tag{20}$$

where t_+ is the positive part of t, namely, $t_+ = \max(0, t)$.

For $\mathbf{E}$: We can obtain the optimization of $\mathbf{E}$ with fixed $\mathbf{w}, \mathbf{Z}$, and $\mathbf{W}$ as follows:

$$\begin{aligned} \mathbf{E}_{t+1} &= \arg\min_{\mathbf{E}} \frac{\mu}{2}\left\|\hat{\mathbf{W}}_t^T\mathbf{X} - \hat{\mathbf{W}}_t^T\mathbf{XZ}_{t+1} - \mathbf{E}\right\|_F^2 \\ &\quad + \left\langle \mathbf{Y}_{1,t}, \hat{\mathbf{W}}_t^T\mathbf{X} - \hat{\mathbf{W}}_t^T\mathbf{XZ}_{t+1} - \mathbf{E}\right\rangle + \lambda\|\mathbf{E}\|_1 \\ &= \arg\min_{\mathbf{E}} \frac{\lambda}{\mu_t}\|\mathbf{E}\|_1 + \frac{1}{2}\left\|\mathbf{E} - \hat{\mathbf{W}}_t^T\mathbf{U}_{t+1} - \mathbf{Y}_{1,t}/\mu_t\right\|_F^2, \end{aligned} \tag{21}$$

where $\mathbf{U}_{t+1} = \mathbf{X} - \mathbf{XZ}_{t+1}$ is defined for simplicity. The optimization of Eq. (21) can be solved by using the shrinkage operator [49].

For $\mathbf{w}$: We can optimize $\mathbf{w}$ with fixed $\mathbf{E}, \mathbf{Z}$, and $\mathbf{W}$ as follows:

$$\begin{aligned} \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} & \frac{1}{2}\left\|\mathbf{y} - (\mathbf{Z}_{t+1}\mathbf{M})^T\mathbf{w}\right\|_2^2 \\ & + \phi\mathbf{w}^T\mathbf{ZLZ}^T\mathbf{w} + \frac{\alpha}{\beta}\|\mathbf{w}\|_2^2. \end{aligned} \tag{22}$$

The above problem is actually the well-known ridge regression, whose optimal solution is $\mathbf{w}_{t+1} = (\mathbf{Z}_{t+1}\mathbf{M}\mathbf{M}^T \mathbf{Z}_{t+1}^T + 2\phi\mathbf{Z}_{t+1}\mathbf{L}\mathbf{Z}_{t+1}^T + \frac{2\alpha}{\beta}\mathbf{I})^{-1}\mathbf{Z}_{t+1}\mathbf{M}\mathbf{y}$.

For $\mathbf{W}$: By setting the derivative of L regarding $\mathbf{W}$ to zero, we have

$$\begin{aligned}&\hat{\mathbf{S}}\hat{\mathbf{W}}_{t+1}(\mathbf{Y}_{2,t} + \mathbf{Y}_{2,t}^T) - 2\delta\beta\mathbf{S}\hat{\mathbf{W}}_{t+1}\\ &+ \mu_t\mathbf{U}_{t+1}\mathbf{U}_{t+1}^T\hat{\mathbf{W}}_{t+1} = \mathbf{U}_{t+1}\mathbf{E}_{t+1}^T - \mathbf{U}_{t+1}\mathbf{Y}_{1,t}^T.\end{aligned} \tag{23}$$

Then, $\mathbf{W}_{t+1}$ can be optimized by solving the Lyapunov equation.

Moreover, the Lagrange multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are updated by the following scheme

$$\begin{aligned}\mathbf{Y}_{1,t+1} &= \mathbf{Y}_{1,t} + \mu_t(\hat{\mathbf{W}}_{t+1}^T\mathbf{U}_{t+1} - \mathbf{E}_{t+1})\\ \mathbf{Y}_{2,t+1} &= \mathbf{Y}_{2,t} + \mu_t(\hat{\mathbf{W}}_{t+1}^T\hat{\mathbf{S}}\hat{\mathbf{W}}_{t+1} - \mathbf{I}).\end{aligned} \tag{24}$$

Algorithm 1. Optimization of Our Proposed Algorithm

```
Input: Feature matrices $\mathbf{X}$, popularity score vector $\mathbf{y}$,
       parameter variables $\lambda, \alpha, \beta, \delta$.
Initialize: $\mathbf{Z}_0 = \mathbf{E}_0 = \mathbf{Y}_{1,0} = \mathbf{Y}_{2,0} = \mathbf{w} = \mathbf{0}, t = 0,$
       $\phi = 1.3, \mu_0 = 10^{-6}, \mu_{max} = 10^6, t_{max} = 10^3.$
1. Compute the covariance matrix $\mathbf{S}_{ij}$ by $\mathbf{S}_{ij} = \mathbf{X}_i\mathbf{X}_j^T$;
2. Initialize $\hat{\mathbf{W}}_0$ by $\hat{\mathbf{W}}_0 = \arg\max_{\hat{\mathbf{W}}} \frac{tr(\hat{\mathbf{W}}^T(\mathbf{S} - \hat{\mathbf{S}})\hat{\mathbf{W}})}{tr(\hat{\mathbf{W}}^T\hat{\mathbf{S}}\hat{\mathbf{W}})}$;
While not converged do
    3. Fix others and update $\mathbf{Z}_{t+1}$:
       $\mathbf{Z}_{t+1} = \arg\min_{\mathbf{Z}} \frac{1}{\tau\mu}\|\mathbf{Z}\|_* + \frac{1}{2}\|\mathbf{Z} - \mathbf{Z}_t + \triangledown_{\mathbf{Z}} h\|_F^2$;
    4. Fix others and update $\mathbf{E}_{t+1}$:
       $\mathbf{E}_{t+1} = \arg\min_{\mathbf{E}} \frac{\lambda}{\mu_t}\|\mathbf{E}\|_1$
           $+ \frac{1}{2}\|\mathbf{E} - \hat{\mathbf{W}}_t^T\mathbf{X} + \hat{\mathbf{W}}_t^T\mathbf{X}\mathbf{Z}_{t+1} - \mathbf{Y}_{1,t}/\mu_t\|_F^2$;
    5. Fix others and update $\mathbf{w}_{t+1}$:
       $\mathbf{w}_{t+1} =$
         $\left(\mathbf{Z}_{t+1}\mathbf{M}\mathbf{M}^T\mathbf{Z}_{t+1}^T + 2\phi\mathbf{Z}_{t+1}\mathbf{L}\mathbf{Z}_{t+1}^T + \frac{2\alpha}{\beta}\mathbf{I}\right)^{-1}\mathbf{Z}_{t+1}\mathbf{M}\mathbf{y}$;
    6. Fix others and update $\hat{\mathbf{W}}_{t+1}$:
       $\hat{\mathbf{S}}\hat{\mathbf{W}}_{t+1}(\mathbf{Y}_{2,t} + \mathbf{Y}_{2,t}^T) - 2\delta\beta\mathbf{S}\hat{\mathbf{W}}_{t+1} + \mu_t\mathbf{U}\mathbf{U}^T\hat{\mathbf{W}}_{t+1}$
        $= 2\mathbf{U}\mathbf{E}_{t+1}^T - \mathbf{U}\mathbf{Y}_1^T,\ \hat{\mathbf{W}}_{t+1} \leftarrow \text{Orthogonal}(\hat{\mathbf{W}}_{t+1})$;
    7. Update the multipliers $\mathbf{Y}_{1,t+1}$ and $\mathbf{Y}_{2,t+1}$:
       $\mathbf{Y}_{1,t+1} = \mathbf{Y}_{1,t} + \mu_t(\hat{\mathbf{W}}_{t+1}^T\mathbf{X} - \hat{\mathbf{W}}_{t+1}^T\mathbf{X}\mathbf{Z}_{t+1} - \mathbf{E}_{t+1})$;
       $\mathbf{Y}_{2,t+1} = \mathbf{Y}_{2,t} + \mu_t(\hat{\mathbf{W}}_{t+1}^T\hat{\mathbf{S}}\hat{\mathbf{W}}_{t+1} - \mathbf{I})$;
    8. Update the parameter $\mu_{t+1}$ by $\mu_{t+1} = \min(\phi\mu_t, \mu_{max})$;
    9. Check the convergence conditions;
End while
Output: $\hat{\mathbf{W}}, \mathbf{E}, \mathbf{Z}, \mathbf{w}$
```

4 Experiments and Results

In this section, we evaluate the effectiveness of our proposed approach on a publicly available micro-video dataset [3]. In Section 4.1, we first briefly describe the micro-video dataset and the experimental settings. We then present various types of extracted features that represent the popularity of micro-videos in Section 4.2 and describe evaluation metrics in Section 4.3. Finally, we provide the experimental results and discussions in Section 4.4.

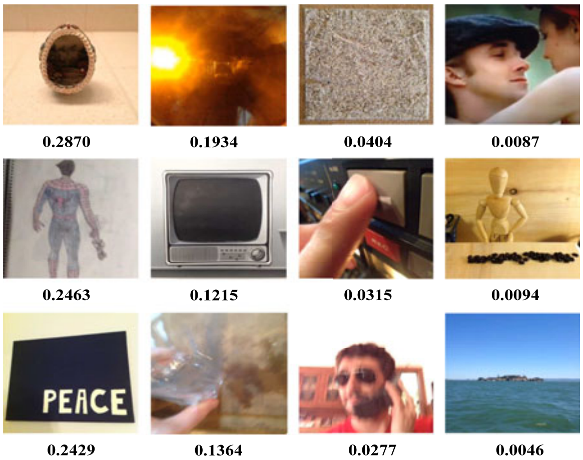


Fig. 2. Sample micro-videos with various popularity scores. The micro-videos are sorted from more popular (left) to less popular (right).

4.1 Micro-Video Dataset

The large-scale micro-video dataset[4] used in this paper was constructed by the Lab for Media Search (LMS) at the National University of Singapore. In total, this dataset contains 303,242 user-generated micro-videos collected from the online micro-video sharing site Vine, which were uploaded by 98,166 users. The length of all micro-videos is no longer than 8 seconds, with approximately 75 percent of the videos being 6-7 seconds. In addition, 120,324 following relationships behind users and 1.6 million video postings from July 2015 to October 2015 are also included in this dataset. Since popularity is highly related to online social interactions, the mean values of four types of statistics, namely, the numbers of comments, reposts, likes and views/loops, are taken into account to formulate the final popularity scores of micro-videos. Fig. 2 shows sample micro-videos that span a wide range of popularity scores.

We tested the prediction performance over 10 random splits of the dataset and report the average results. In each round, we used 90 percent of the micro-videos for training and the remaining for testing. We empirically set the adaptive parameters as $\alpha = 1$, $\delta = 0.1$, and $\lambda = 0.01$ as default. The trade-off parameters β and ϕ in TLRMVR model are selected by a grid-search approach. We first performed a coarse grid. Once we identified a ideal region, we then conducted a finer grid search on that region. Finally, we set $\phi = 0.1$ and $\beta = 0.5$.

4.2 Feature Extraction

In this section, we represent micro-videos using four-view features extracted from visual, acoustic, contextual, and social modalities (i.e., we denote $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ as the feature matrices corresponding to visual, acoustic, textual, and social views, respectively).

4.2.1 Visual Features

Due to the short-length property of micro-videos, topic variety within one micro-video is usually limited, therefore the keyframe-based representation strategy becomes more robust in terms of representing its intrinsic topic. Inspired by this property, the visual features of micro-videos were

[4] http://acmmm2016.wixsite.com/micro-videos.

obtained by adopting the average pooling operation to fuse features extracted from certain keyframes.

- *Color Histogram.* As noted in [38], simple image features show little correlation with popularity prediction. Because a color histogram can easily attract more attention by revealing striking colors, the color space is grouped into 50 distinct colors, resulting in a 50-D vector color histogram feature for each frame.
- *Object Features.* It has been demonstrated that the high-level object representation is an effective feature in popularity prediction. Due to the strong performance of deep convolutional neural networks (CNNs) in visual understanding tasks [50], the well-trained "AlexNet" ImageNet model is applied to directly represent keyframes. The output of the last fully connected layer fc7 is taken as the input to a 1,000-way softmax, and a distribution over the 1,000 class labels is produced, which is treated as the final representation in terms of object features.
- *SentiBank Features.* Some studies have been conducted to investigate the influence of sentiment on analysing multimedia content [31], [51]. For example, Gelli et al. [31] performed experiments on large-scale datasets, suggesting that sentiment concepts, i.e., adjective-noun pairs (ANPs), have a positive impact on popularity prediction. Chen et al. [52] trained a deep CNN model called DeepSentiBank for the classification of visual sentiment concepts, in which 2,089 ANPs such as "cut dog" are trained with 867,919 images. The output of the last fully connected layer is taken as the input to a 2,089-way softmax, and a distribution over the 2,089 ANPs is generated as high-level sentiment features.
- *Aesthetic Features.* Aesthetics specify the highly subjective nature of human perception. Studies of aesthetics [53], [54] show that a high aesthetic quality makes some images more appealing than others. Following the video aesthetic assessment by Bhattacharya et al. [55], 149-D visual statistical features are extracted to describe micro-videos at the frame level, including dark channel, sharpness, eye sensitivity, low depth of field, white balance, colorfulness, color harmony, and color harmony statistics.

Before analyzing the data, we first normalized each type of textual features, respectively. We then concatenated all types of features to form a 3,288-D textual feature representation. Finally, all feature vectors are normalized to unit $\mathcal{L}_2$-norm length.

4.2.2 Acoustic Features

Acoustic features are essential to various tasks, cross-media correlation, for example. Acoustic information can provide complementary cues to the visual content, particularly for insufficient visual information in videos. Previous research in micro-videos also employed acoustic information embedded in micro-videos to improve learning performance [2], [3]. For example, Chen et al. [3] took acoustic features as an input and investigated the influence of acoustic features on micro-video popularity prediction. Zhang et al. [2] extracted acoustic features based on a stack denoising autoencoder for a more comprehensive representation of micro-videos. Following the setup of Chen et al. [3], we used 36-D features extracted from the audio channel to represent the acoustic modality of micro-videos, including mel-frequency cepstral coefficients (MFCCs), energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid, and spectral flux. The ranges of the acoustic feature are continuous real values. We normalized all values to unit $\mathcal{L}_2$-norm length.

4.2.3 Textual Features

Additional textual information provides new opportunities for understanding micro-video content from different aspects. Recent work in popularity prediction of social media has considered textual information as an indispensable component to improve the prediction. In some cases, the textual descriptions associated with micro-videos incorporate topic information and sentiment of the publisher, which are critical to popularity prediction. For example, Mishne and Glance [56] utilized sentiment values of the contexts as an indicator to predict the popularity of movies in terms of sales. Sentence2Vector[5] is a classical textual feature extraction tool, which is utilized to produce 100-D features for topic representation of micro-videos. Stanford CoreNLP tools[6] provides a tool for the sentiment analysis of texts. By leveraging the sentiment analysis tool, each micro-video is assigned to a sentiment score, which is an integer ranging from 0 to 4 corresponding to "very negative", "negative", "neutral", "positive", and "very positive", respectively. Before analyzing the data, we first normalized each type of textual features repsectviely and concatenated them together. Then, all feature values are normalized to unit $\mathcal{L}_2$-norm length.

4.2.4 Social Features

Although some related works have demonstrated that low-level visual features and high-level semantic features are able to predict popularity to some extent, yet social cue is a significant factor in determining how widely a micro-video is spreading. For example, micro-video followees, particularly popular followees, often act as influential leaders due to their noticeable impact on the followers' later decisions. Thus, a micro-video uploaded by a popular followee tends to attract more potential audiences. Thus, 4 types of social cues are encoded to characterize the popularity of a micro-video uploader:

- *Followee-Follower Count*. The number of followers and followees for a given publisher.
- *Loop Count*. The total number of loops of a micro-video after it is uploaded.
- *Post Count*. The number of posts per publisher.
- *Twitter Verification*. A binary value reflecting whether the publisher is a verified user.

All social features vectors are normalized to unit $\mathcal{L}_2$-norm length.

4.3 Evaluation Metric

We adopted the typical normalized Mean Squared Error (nMSE) [57] to measure the consistency between predictions

[5] https://github.com/klb3713/sentence2vec.

[6] http://stanfordnlp.github.io/CoreNLP/.

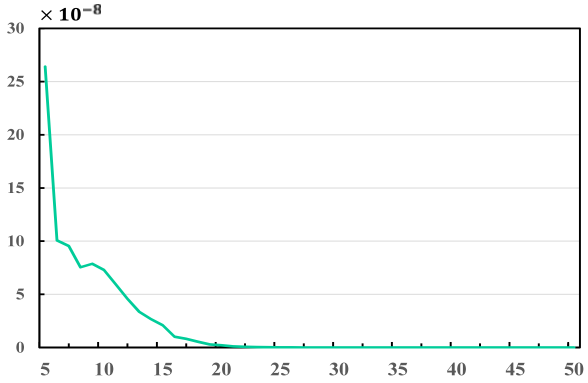


Fig. 3. The convergence curve of our proposed TLRMVR method. The horizontal axis represents the number of iterations, and the vertical axis is the divergence between two consecutive measured **Z**s.

and ground truths. We also performed a pair-wise t-test analysis between the proposed TLRMVR method and each of the other methods based upon the 10-round results. The nMSE value between the predicted results and the ground-truth is defined as

$$nMSE = \frac{1}{M\sigma^2}\sum_{i=1}^{M}(v_i - s_i)^2, \tag{25}$$

where v_i and s_i is the predicted popularity score and the real score of the ith micro-video; σ is the standard deviation of the ground-truths. A lower nMSE value indicates better performance.

4.4 Results and Discussions

To comprehensively validate the proposed algorithm, in the following experiments, we justified the proposed algorithm from the following six perspectives:

- *Convergence analysis*. We tested the convergence of our algorithms based on the proposed alternating algorithm.
- *Component analysis*. To verify the effectiveness of different components in our proposed scheme, we compared the prediction performance by removing each component in our method.
- *Feature analysis*. To evaluate how features contribute to the micro-video popularity prediction, we considered two forms of evaluation: i) performance comparison among different views and ii) performance comparison among different visual features.
- *Parameter sensitivity analysis*. We conducted experiments to investigate the influence of various weighting parameters on the prediction accuracy.
- *Comparison with state-of-the-art methods*. Performance comparisons with several state-of-the-art algorithms were conducted to demonstrate the effectiveness of our method.

4.4.1 Convergence Analysis

In this section, we tested the convergence of our objective function based on the proposed alternating algorithm and randomly selected a trial to report the results. Because **Z** is used for predicting the popularity of micro-videos, we would like to measure the variance between two sequential **Z**s by the following metric.

$$D(t) = \|\mathbf{Z}_t - \mathbf{Z}_{t-1}\|_F. \tag{26}$$

This will guarantee that the final feature results will not be drastically changed. Fig. 3 presents the absolute values of the variance during the iterations. As shown in this figure, the divergence values obtained for our proposed algorithm decrease rapidly with increasing numbers of iterations and converge after approximately 20 iterations. Based on the above analysis, the iterative criteria are essential to guarantee the convergence of our objective function. Therefore, in this paper, we used the relative change between two consecutive iterations falling below a threshold of 1e-3 and a maximum of 30 iterations as the stopping criteria for our proposed method.

TABLE 1
Performance Comparison of Involved Components in Our Proposed Framework

	noLR	noGR	noMR	noSP	TLRMVR
Top50	0.296	0.347	0.326	0.204	0.309
Top100	0.291	0.317	0.311	0.201	0.280
Top200	0.276	0.296	0.285	0.192	0.276
Bottom200	0.253	0.271	0.269	0.172	0.265
Bottom100	0.249	0.258	0.254	0.161	0.256
Bottom50	0.246	0.251	0.251	0.157	0.249
nMSE	0.950	0.949	0.949	0.973	**0.934**
P-value	<0.05	<0.05	<0.05	<0.05	-

4.4.2 Component Analysis

To validate the contributions of each component in our proposed framework, we compared the prediction performance by removing the relevant components:

- *noLR*. We eliminated the influence of the low-rank constraint imposed on **Z** by replacing it with the Frobenius norm.
- *noGR*. We eliminated the influence of the graph regularization term by setting $\phi = 0$.
- *noMR*. We eliminated the effect of multi-view embedding learning by setting $\delta = 0$.
- *noSP*. We eliminated the influence of supervised information, i.e., view information and regression information, by discarding both the regression coefficient and view-specific transformation matrices learning.

In the case of noSP, our algorithm degenerates to a typical unsupervised low-rank feature representation. In order to get comparable results, a least squares regression model is trained to predict popularity scores. Table 1 shows the prediction results of different schemes. In this table, we selected the top $50, 100, 200$ images with the highest ground truth and the bottom $50, 100, 200$ images with the lowest ground truth to report the average popularity scores based on their predicted results. As shown in this table, the predicted popularity scores over different ranges are Top50 $>$ Top100 $>$ Top200 $>$ Bottom200 $>$ Bottom100 $>$ Bottom50, illustrating that the behavior of the predicted results is reasonable. Moreover, we sorted the nMSE values of different

TABLE 2
Performance Comparison with Different Visual-Level Feature Combinations at Predicting Micro-Video Popularity

	Color	Object	Sentiment	Aesthetics	All
Top50	0.364	0.231	0.247	0.203	0.309
Top100	0.325	0.229	0.231	0.194	0.280
Top200	0.301	0.193	0.204	0.174	0.276
Bottom200	0.279	0.184	0.199	0.167	0.265
Bottom100	0.254	0.182	0.193	0.164	0.256
Bottom50	0.253	0.177	0.191	0.160	0.249
nMSE	0.975	0.967	0.969	0.971	**0.934**
P-value	<0.05	<0.05	<0.05	<0.05	-

TABLE 3
Performance Comparison with Different View-Level Feature Combinations at Predicting Micro-Video Popularity

	T+V+A	T+A+S	T+V+S	V+A+S	TLRMVR
Top50	0.273	0.241	0.289	0.272	0.309
Top100	0.241	0.201	0.250	0.227	0.280
Top200	0.238	0.255	0.249	0.225	0.276
Bottom200	0.233	0.199	0.247	0.218	0.265
Bottom100	0.224	0.179	0.229	0.213	0.256
Bottom50	0.218	0.172	0.221	0.201	0.249
nMSE	0.979	0.970	0.958	0.955	**0.934**
P-value	<0.05	<0.05	<0.05	<0.05	-

methods in descending order and found that noSP $>$ noLR $>$ noGR $=$ noMR; thus, the following conclusions are obtained: 1) Without supervised information, noSP performs the worst, indicating that the valuable supervised information is essential to learn a more robust prediction model. Moreover, noSP separates micro-video plurality prediction into two phases, which may lead to sub-optimal prediction results. 2) noMR and noLR impose similar significant effects on the prediction results, which means the low-rank representation and multi-view embedding learning are important in reducing the heterogeneous gap among features and alleviating the influence of feature noises. 3) Our proposed TLRMVR outperforms noGR, which demonstrates that our proposed method benefits from the use of graph regularization. This result further indicates that multi-graph regularization can indeed be employed to address multi-view feature fusion problem. 4) P-value [58] is adopted to assess whether the superiority of the TLRMVR method is statistically significant. We can discover that the P-values are smaller than the significance level of 0.05, which indicates that the null hypothesis is clearly rejected and that the improvements of TLRMVR are statistically significant.

4.4.3 Feature Analysis

Under our proposed framework, we investigated the influence of different features on the micro-video popularity prediction from two perspectives: i) performance comparison of different visual-level feature combinations and ii) performance comparison of different view-level feature combinations.

We first selected one of four visual features to represent the visual content of micro-videos and integrate with contextual, social, and acoustic cues together to conduct our experiments. Table 2 reports the average results over 10 random splits in terms of nMSE and P-value. From Table 2, we can observe the following results: 1) Object features perform the best among visual features, indicating that object semantics can encode important information that makes a micro-video popular. 2) Visual sentiment has a significant influence on prediction performance, illustrating that high-level sentiment semantics are helpful for micro-video popularity prediction. 3) The aesthetics exhibits better performance than the color histogram since aesthetic features specify the highly subjective nature of human perception. 4) The worst performance is still achieved by color histogram, although color histogram is effective in modeling the color perception of the human visual system. 5) The best performance is achieved when all visual features are combined, illustrating the benefit of exploiting the complementary information offered by different visual representations.

Subsequently, we evaluated how various view-level feature combinations contribute to the popularity of micro-videos under our proposed framework. For simplicity, the features extracted from textual, visual, acoustic, and social cues are indicated as "T", "V", "A", and "S", respectively. Table 3 shows the average results in terms of nMSE and P-value. From Table 3, we can observe the following results: 1) Similar to other existing works, "T+V+A" provides the most unsatisfactory results when removing social cues, which indicates that social cues can largely facilitate popularity prediction compared to other types of cues. 2) The prediction performance of "T+A+S" sharply decreases after removing visual cues. This result shows that visual cues of micro-videos serve as an indispensable component to further improve the prediction performance. 3) "V+A+S" yields a good result of nMSE = 0.955 compared to the other forms of combinations, indicating that textual cues exhibit little effect on popularity. One possible reason causing this phenomenon is that there are quite a fair number of micro-videos that lack textual descriptions. Moreover, the weak correlation between textual descriptions and micro-videos is also a common cause of this effect. 4) When combined all view features together, the best performance is achieved with a minimum nMSE of 0.934. Additionally, it could therefore be concluded that the sequences of all cues, which are sorted in descending order in terms of their importance, is social $>$ visual $>$ acoustic $>$ textual cues.

4.4.4 Parameter Sensitivity Analysis

Among all the parameters in our proposed objective function, we found that the parameters ϕ and β play significant roles in affecting the prediction results. As shown in Eq. (22), the trade-off parameter ϕ is used to balance the effects between the graph regularization and ridge regression and the trade-off parameter β is mainly used to control the effect of the supervised loss term. Therefore, we would like to evaluate different values of ϕ and β to investigate the variation in prediction performance. In this experiment, the parameter ϕ and β are selected via a grid search in a heuristic manner, ranging from 0.05 to 0.30 with an interval 0.05 and ranging from 0.25 to 1.25 with an interval 0.25, respectively. nMSE results for various values of ϕ and β are reported in Tables 4 and 5, respectively. As shown in this table, the best performance is achieved when $\phi = 0.10$ and $\beta = 0.50$. In fact, when ϕ is set 0, our proposed method is

TABLE 4
Performance Comparison with Different ϕ on Our Proposed Framework

	0.05	0.10	0.15	0.20	0.25	0.30
Top50	0.370	0.309	0.283	0.238	0.230	0.198
Top100	0.347	0.280	0.269	0.227	0.219	0.187
Top200	0.330	0.276	0.251	0.212	0.205	0.175
Bottom200	0.309	0.265	0.241	0.204	0.197	0.168
Bottom100	0.298	0.256	0.231	0.196	0.189	0.162
Bottom50	0.294	0.249	0.227	0.193	0.186	0.159
nMSE	0.948	**0.934**	0.953	0.957	0.958	0.961
P-value	<0.05	-	<0.05	<0.05	<0.05	<0.05

TABLE 6
Performance Comparison with Different Reduced Dimensions D on Our Proposed Framework

	10	20	30	40	50	60
Top50	0.319	0.308	0.318	0.316	0.316	0.297
Top100	0.288	0.279	0.286	0.277	0.277	0.270
Top200	0.274	0.276	0.275	0.274	0.274	0.262
Bottom200	0.269	0.265	0.269	0.267	0.267	0.256
Bottom100	0.250	0.256	0.252	0.249	0.249	0.245
Bottom50	0.243	0.249	0.243	0.241	0.241	0.236
nMSE	0.950	**0.934**	0.947	0.949	0.951	0.953
P-value	<0.05	-	<0.05	<0.05	<0.05	<0.05

reduced to discard the graph regularization term, which easily induces the overfitting problem. If β is set 0, our proposed method is equivalent to discard the supervised information and easily induces unsatisfactory results. This conclusions can be verified in Section 4.4.2.

We also evaluated the influence of various dimensions of the projection matrices. The performance of TLRMVR with different D from 10 to 60 is illustrated in Table 6. From the table, we discovered that the best dimension is 20. Too small or too large a dimension leads to a suboptimal prediction performance. It is a reasonable choice to take 20 as the reduced dimension in consideration of the complementary properties of different views.

4.4.5 Comparison with State-of-the-Art Methods

We compared our proposed scheme with several existing state-of-the-art methods, including multiple linear regression, lasso regression, support vector regression [59], RegMVMT [60], multi-feature learning via hierarchical regression (MLHR) [61], multiple social network learning (MSNL) [62], multi-view discriminant analysis [63], transductive multi-modal learning [3], and extreme learning machine (ELM) [64].

- *MLR*. Multiple linear regression attempts to capture the dependency between two or more independent variables and a response variable using a linear equation, which is an extension of classical linear regression.
- *Lasso*. Lasso regression considers both variable selection and regularization to enhance the prediction performance.
- *SVR*. Support vector regression [59] is a classical regression technique with a maximum margin criterion. We combined all the features together with an RBF kernel to learn a non-linear SVR in a high-dimensional kernel-induced feature space.

TABLE 5
Performance Comparison with Different β on Our Proposed Framework

	0.25	0.50	0.75	1	1.25
Top50	0.309	0.308	0.322	0.200	0.204
Top100	0.305	0.279	0.294	0.185	0.189
Top200	0.285	0.276	0.283	0.181	0.186
Bottom200	0.263	0.265	0.268	0.175	0.181
Bottom100	0.257	0.256	0.257	0.170	0.176
Bottom50	0.252	0.249	0.251	0.166	0.172
nMSE	0.949	**0.934**	0.950	0.962	0.968
P-value	<0.05	-	<0.05	<0.05	<0.05

- *RegMVMT*. RegMVMT [60] is an inductive learning framework to address the general multi-view learning problem, in which the co-regularization technique is utilized to enforce the agreement with other views on unlabeled samples.
- *MLHR*. The multi-feature fusion via hierarchical regression [61] is a semi-supervised learning method, which has been developed to explore the structural information embedded in data from the view of multi-feature fusion.
- *MSNL*. Multiple social network learning [62] is proposed to address the incomplete data in source confidence and source consistency by modeling source confidence and source consistency simultaneously.
- *MvDA*. Multi-view discriminant analysis (MvDA) [63] is a multi-view learning model, which has been developed to search for a latent common space by enforcing the view-consistency of multi-linear transforms.
- *TMALL*. The transductive multi-modal learning [3] model is presented for predicting the popularity of micro-videos, in which different modal features can be unified and preserved in a latent common space to address the insufficient information problems.
- *ELM*. As ELM [65], [66] can embed a wide type of feature mappings, Huang et al. [64] extended ELM to kernel learning and proposed a unified learning mechanism for regression applications with higher scalability and less computational complexity.

Table 7 reports the prediction performances of our proposed method and other state-of-the-art algorithms. From this table, we have the following observations. 1) Our proposed TLRMVR performs the best among all the comparative methods. 2) Lasso and MLR performs the worst, as expected, indicating that simple feature selection and linear regression are insufficient to predict the popularity of micro-videos. 3) In contrast to Lasso and MLR, the algorithms, including RegMVMT, MLHR, MSNL, MvDA, and TMALL, also performs comparably, which can be attributed to their ability to solve the multi-view/modal feature fusion problem. 4) After employing the RBF kernel to deal with multiple features, the SVR model provides a significant improvement in the micro-video popularity prediction tasks. 5) As stated in [64], SVR provides a suboptimal learning solution compared to ELM. Accordingly, the results

TABLE 7
Performance Comparison Between Our Proposed Method and Several State-of-the-Art Methods

Methods	nMSE	P-value
MLR	1.442 ± 2.55E-01	1.05E-07
Lasso	1.568 ± 1.72E-01	4.42E-08
SVR	0.991 ± 5.00E-02	7.36E-06
RegMVMT	1.058 ± 4.33E-05	1.88E-03
MLHR	1.167 ± 1.40E-02	4.75E-06
MSNL	1.098 ± 1.30E-01	2.11E-04
MvDA	0.982 ± 7.00E-03	2.62E-05
TMALL	0.979 ± 9.42E-03	1.43E-08
ELM	0.982±6.68E-05	3.71E-07
TLRMVR	**0.934 ± 7.67E-04**	-

present that ELM achieves better prediction performance than SVR. 6) Although MSNL and TMALL are appropriate to deal with incomplete data, TLRMVR still outperforms them, thus demonstrating the effectiveness of our approach.

4.4.6 Complexity Discussion

In order to analyze the complexity of TLRMVR, we suppose that the number of samples is larger than the dimension of data, i.e., $(N + M) > (D_1 + D_2 + \cdots + D_K)$. As discussed in previous sections, we can find that the main computational complexity comes from the following parts:

- Nuclear norm calculation in step 3.
- Matrix inverse calculation in step 5.
- Solving the Lyapunov equation in step 6.

The computational complexity of Nuclear norm is at most $O((N + M)^3)$. The matrix inverse costs $O((N + M)^3)$. The typical cost of the Lyapunov equation needs $O((N + M)^3)$. If the algorithm converges within T iteration steps for its outer loop, the upper bound of the complexity is $O(3T(N + M)^3)$. The simulations of our proposed algorithm are carried out in MATLAB 7.0.1 environment running in Core 3 Quad, 3.6-GHZ CPU with 8-GB RAM. The learning and testing processes over all micro-videos can be accomplished within 1,627 seconds. The speed bottleneck lies in the number of samples. Therefore, to handle large-scale dataset, Coppersmith and Winograd [67] presented a new method to accelerate matrix inversion to $O((N + M)^{2.376})$. Liu et al. [68] offered a more efficient method to solve Nuclear norm calculation.

5 Conclusion and Future Work

In this paper, we have proposed a novel low-rank multi-view embedding framework to alleviate the heterogeneous, interconnected, and noisy problems in micro-video popularity prediction. By taking advantages of low-rank representation and multi-view learning, we effectively integrated all heterogeneous features extracted from different views into a common feature subspace and achieved enhanced robust feature representation for regression analysis. We also designed an effective optimization algorithm to solve the proposed model. Experimental results on a publicly available dataset demonstrated that the performance of our proposed scheme obtained superior performance over state-of-the-art methods.

In future, we will seek a more flexible mechanism for micro-video popularity prediction in which features fusion can be performed by a hierarchical strategy. To cope with the rapid increase in data size, we will attempt to build a scalable learning framework by exploring the underlying structure of the whole dataset with both data points and anchors [69]. Besides, motivated by recent progress in ELM [64] [70] in dealing with regression and dimension reduction problems, we will also consider advantages of ELM in the wide variety of feature mapping functions and the ability of handling large data to address micro-video popularity prediction.

References

[1] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes, "6 seconds of sound and vision: Creativity in micro-videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 4272–4279.

[2] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1415–1424.

[3] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, "Micro tells macro: Predicting the popularity of micro-videos via a transductive model," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 898–907.

[4] P. X. Nguyen, G. Rogez, C. Fowlkes, and D. Ramanan, "The open world of micro-videos," arXiv:1603.09439, 2016.

[5] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[6] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1615–1622.

[7] X. Zhang, F. Sun, G. Liu, and Y. Ma, "Fast low-rank subspace segmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1293–1297, May 2014.

[8] S. Li and Y. Fu, "Learning balanced and unbalanced graphs via low-rank coding," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1274–1287, May 2015.

[9] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2429–2443, Jun. 2016.

[10] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900–1913, Aug. 2016.

[11] Z. Zhang, S. Yan, and M. Zhao, "Similarity preserving low-rank representation for enhanced data representation and effective subspace learning," *Neural Netw.*, vol. 53, pp. 81–94, 2014.

[12] Z. Zhang, M. Zhao, F. Li, L. Zhang, and S. Yan, "Robust alternating low-rank representation by joint LP-and l2, P-norm minimization," *Neural Netw.*, vol. 96, pp. 55–70, 2017.

[13] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by nuclear/l2, 1-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, Apr. 2017.

[14] J. Zhuang, T. Mei, S. C. Hoi, X.-S. Hua, and Y. Zhang, "Community discovery from social media by low-rank matrix recovery," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, pp. 67:1–19, 2015.

[15] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 1192–1198.

[16] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley Online Library, 2002.

[17] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recong.*, 2012, pp. 2618–2625.

[18] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 676–683.

[19] P. Zhou, Z. Lin, and C. Zhang, "Integrated low-rank-based discriminative feature learning for recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1080–1093, May 2016.

[20] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 2149–2155.

[21] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2778–2784.

[22] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 110–119.

[23] Z. Ding and Y. Fu, "Robust multi-view subspace learning through dual low-rank decompositions," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1181–1187.

[24] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, 2017.

[25] S. Xiang, Y. Zhu, X. Shen, and J. Ye, "Optimal exact least squares rank minimization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 480–488.

[26] S. Zheng, X. Cai, C. H. Ding, F. Nie, and H. Huang, "A closed form solution to multi-view low-rank regression," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1973–1979.

[27] Y. Bae and H. Lee, "Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 63, no. 12, pp. 2521–2535, 2012.

[28] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in twitter," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 64, no. 7, pp. 1399–1410, 2013.

[29] P. J. McParlane, Y. Moshfeghi, and J. M. Jose, "Nobody comes here anymore, it's too crowded; predicting image popularity on flickr," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, pp. 385–391.

[30] S. Cappallo, T. Mensink, and C. G. Snoek, "Latent factors of visual popularity prediction," in *Proc. Int. Conf. Multimedia Retrieval*, 2015, pp. 195–202.

[31] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 907–910.

[32] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.

[33] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu, "On popularity prediction of videos shared in online social networks," in *Proc. ACM Int. Conf. Inform. Knowl. Manag.*, 2013, pp. 169–178.

[34] J. Liu, Y. Yang, Z. Huang, and Y. Yang, "On the influence propagation of web videos," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1961–1973, Aug. 2014.

[35] T. Trzcinski and P. Rokita, "Predicting popularity of online videos using support vector regression," arXiv:1510.06223, 2015.

[36] M. Vasconcelos, J. M. Almeida, and M. A. Gonçalves, "Predicting the popularity of micro-reviews: A foursquare case study," *Inform. Sci.*, vol. 325, pp. 355–374, 2015.

[37] B. Wu and H. Shen, "Analyzing and predicting news popularity on twitter," *Int. J. Inform. Manag.*, vol. 35, no. 6, pp. 702–711, 2015.

[38] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proc. ACM Int. Conf. World Wide Web*, 2014, pp. 867–876.

[39] L. C. Totti, F. A. Costa, S. Avila, E. Valle, W. Meira, and V. Almeida, "The impact of visual attributes on online image diffusion," in *Proc. ACM Web Sci. Conf.*, 2014, pp. 42–51.

[40] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1882–1895, 2016.

[41] W. Ding, Y. Shang, L. Guo, X. Hu, R. Yan, and T. He, "Video popularity prediction by sentiment propagation via implicit network," in *Proc. ACM Int. Conf. Inform. Knowl. Manag.*, 2015, pp. 1621–1630.

[42] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1255–1267, Jun. 2013.

[43] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[44] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conf. Data Mining Data Warehouses*, 2010, pp. 1–4.

[45] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[46] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[47] H. Peng, K. Li, B. Li, H. Ling, W. Xiong, and W. Hu, "Predicting image memorability by multi-view adaptive regression," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1147–1150.

[48] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[49] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," Electric. Comput. Eng. Dept., Univ. Illinois at Urbana-Champaign, Champaign, IL, USA, Tech. Rep., UILU-ENG-09–2215, 2010.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[51] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1050–1062, May 2016.

[52] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," arXiv:1410.8586, 2014.

[53] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, Jan. 2014.

[54] L. Nie, R. Hong, L. Zhang, Y. Xia, D. Tao, and N. Sebe, "Perceptual attributes optimization for multivideo summarization," *IEEE Trans. Cybernatrics*, vol. 46, no. 12, pp. 2991–3003, Dec. 2016.

[55] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model foraesthetic assessment of videos," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 361–364.

[56] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment," in *Proc. AAAI Spring Symp.: Comput. Approaches Anal. Weblogs*, 2006, pp. 155–158.

[57] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua, "Beyond doctors: Future health prediction from multimedia and multimodal observations," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 591–600.

[58] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 396–409, Feb. 2015.

[59] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[60] J. Zhang and J. Huan, "Inductive multi-task learning with multiple view data," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 543–551.

[61] Y. Yang, J. Song, Z. Huang, and Z. Ma, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 572–581, Mar. 2013.

[62] X. Song, L. Nie, L. Zhang, M. Akbari, and T.-S. Chua, "Multiple social network learning and its application in volunteerism tendency prediction," in *Proc. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2015, pp. 213–222.

[63] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[64] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[65] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2004, vol. 2, pp. 985–990.

[66] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

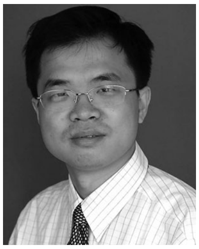
[67] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," in *Proc. ACM Symp. Theory Comput.*, 1987, pp. 1–6.

[68] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[69] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1864–1877, Jul. 2016.

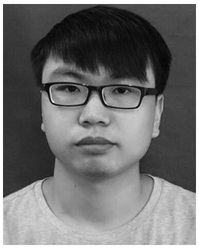
[70] L. L. C. Kasun, Y. Yang, G.-B. Huang, and Z. Zhang, "Dimension reduction with extreme learning machine," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3906–3918, Aug. 2016.

Peiguang Jing received the MS degree in signal and information processing from Tianjin University, Tianjin, China, in 2012, where he is currently working toward the PhD degree in the School of Electrical and Information Engineering. He was a visiting student with the National University of Singapore, from 2014 to 2015. His current research interests include multimedia content analysis, and tensor decomposition.



Yuting Su received the BS, MS, and the PhD degrees from Tianjin University, Tianjin, China, in 1995, 1998, and 2001 respectively. He is currently a professor in the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include computer vision, multimedia content analysis, information security, and tensor decomposition.

Liqiang Nie received the BE degree from the Xian Jiaotong University of China, Xian, and the PhD degree from the National University of Singapore, in 2009 and 2013, respectively. He is currently a professor in the School of Computer Science and Technology, Shandong University. Prior that, he was a research fellow in the School of Computing, National University of Singapore. His research interests include multimeida computing, information retrieval, and their applications in healthcare analytics. Various parts of his work have been published in top forums, such as the ACM Special Interest Group on Information Retrieval, *ACM Multimedia*, the *ACM Transactions on Information Systems*, and the *IEEE Transactions on Multimedia*. He has served as a reviewer for various journals and conferences. He is a member of the IEEE.



Xu Bai received the BS degree from Tianjin University, Tianjin, China, in 2016, where he is currently working toward the MS degree in the School of Electrical and Information Engineering. His research interest include the multimedia content analysis.

Jing Liu received the BE and PhD degrees from Shanghai Jiao Tong University, Shanghai, China, in 2011 and 2017, respectively. She is currently an assistant professor in the Multimedia Institute at Tianjin University. From 2014 to 2015, she was a visiting student in the Department of Computer Science and Engineering, State University of New York at Buffalo. Her research interests include multimedia signal processing and perceptual visual processing. She is a member of the IEEE.

Meng Wang received the BE and PhD degrees in the special class for the gifted young from the Department of Electronic Engineering and Information Science, the University of Science and Technology of China, Hefei, China, respectively. He is currently a professor with the Hefei University of Technology, Hefei. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He received the Best Paper Awards successively from the 17th and 18th ACM International Conference on Multimedia, the Best Paper Award from the 16th International Multimedia Modeling Conference, and the Best Demo Award from the 20th ACM International Conference on Multimedia. He is a member of the IEEE.