Multi-view Manifold Learning for Media Interestingness Prediction

Yang Liu
[1]Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
[2]IRACE, HKBU, Shenzhen, China
csygliu@comp.hkbu.edu.hk

Zhonglei Gu
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
crossgate9@gmail.com

Yiu-ming Cheung
[1]Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
[2]IRACE, HKBU, Shenzhen, China
ymc@comp.hkbu.edu.hk

Kien A. Hua
Department of Computer Science
University of Central Florida
Orlando, Florida, USA
kienhua@cs.ucf.edu

ABSTRACT

Media interestingness prediction plays an important role in many real-world applications and attracts much research attention recently. In this paper, we aim to investigate this problem from the perspective of supervised feature extraction. Specifically, we design a novel algorithm dubbed Multi-view Manifold Learning (M^2L) to uncover the latent factors that are capable of distinguishing interesting media data from non-interesting ones. By modelling both geometry preserving criterion and discrimination maximization criterion in a unified framework, M^2L learns a common subspace for data from multiple views. The analytical solution of M^2L is obtained by solving a generalized eigen-decomposition problem. Experiments on the Predicting Media Interestingness Dataset validate the effectiveness of the proposed method.

KEYWORDS

Media interestingness analysis; multi-view manifold learning

ACM Reference format:
Yang Liu, Zhonglei Gu, Yiu-ming Cheung, and Kien A. Hua. 2017. Multi-view Manifold Learning for Media Interestingness Prediction. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania,* , 7 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079021

1 INTRODUCTION

Media interestingness analysis, which aims to automatically analyze media data and identify the content which is considered to be interesting for the users, plays an important role in many real-world applications such as image/video retrieval, summarization, and recommendation [3, 11, 16, 19, 39, 46].

With the rapid development of computing resources, computational explorations on media interestingness analysis have received much attention recently. So far, the investigations have primarily focused on images and videos, for the vast amount of rich information these two types of media could provide. Dhar et al. utilized high-level image features, i.e., compositional attributes, content attributes, and sky-illumination attributes, to predict the interestingness of Flickr photos [9]. Gygli et al. presented a computational model that predicts the interestingness of images by measuring three important cues: unusualness, aesthetics, and general preferences [13]. Grabner et al. generated four different cues, i.e., emotion, complexity, novelty, and learning, from the low-level features, and constructed the interestingness predictor based on the combination of these cues [12]. Liu et al. introduced a method to measure the frame interestingness of a travel video by comparing the SIFT features of the frame with those of the web photos [23]. Jiang et al. conducted a pilot study to understand the human perception on video interestingness, and used a computational model to predict the interestingness of videos [16]. Soleymani validated that the intrinsic pleasantness, arousal, visual quality, and coping potential are important factors contributing to visual interest in digital photos and developed a system to detect these attributes from low-level features [39]. In a recently organized *Predicting Media Interestingness Task* [8], various computational models have been utilized to predict the interestingness of given images and videos [6, 22, 35, 37, 41, 44].

Although tremendous strides forward have already been made in analyzing and predicting media interestingness, most of the computational models still utilize manually selected or pre-defined features, which are somewhat subjective and

ICMR '17, June 6–9, 2017, Bucharest, Romania
© 2017 ACM. ACM ISBN 978-1-4503-4701-3/17/06. . . $15.00
DOI: http://dx.doi.org/10.1145/3078971.3079021

might lose important structural and discriminant information. In order to automatically learn the useful and informative features for media interestingness prediction, we formulate the task as a supervised feature extraction problem and propose to utilize manifold learning to solve it, since the distributions of many kinds of media data, such as images and videos, match the manifold assumption very well [32, 34], which makes manifold learning being an appropriate tool for media data analysis. Moreover, the media data can often be characterized by features from multiple views [10, 18]. For instance, images or video frames could be taken from different angles or described by various kinds of visual features such as HoG, SIFT, GIST, etc. In order to explore multi-view media data under the manifold learning framework, we propose a novel algorithm dubbed Multi-view Manifold Learning ($\mathrm{M}^2\mathrm{L}$) to uncover the latent factors that distinguish interesting media data from non-interesting ones. The proposed method aims to preserve both the geometric structure and the interestingness information of the original dataset by mapping the multi-view data to a common subspace, in which the cross-view correlation is taken into consideration.

The rest of the paper is organized as follows. In Section 2, we briefly review some representative works on manifold learning for multimedia data analysis and multi-view dimensionality reduction. The details of the proposed method, $\mathrm{M}^2\mathrm{L}$, is presented in Section 3. In Section 4, we evaluate the performance of $\mathrm{M}^2\mathrm{L}$ on the Predicting Media Interestingness Dataset [8]. The paper is concluded in Section 5.

2 RELATED WORKS

2.1 Manifold Learning for Multimedia Data Analysis

Manifold learning, which uncovers the intrinsic structure of data based on the assumption that the original observations lie on or close to the low-dimensional nonlinear manifold, has received much attention in the past two decades. The most representative manifold learning algorithms include isometric feature mapping (Isomap) [40], locally linear embedding (LLE) [36], and Laplacian eigenmaps (LE) [2]. Isomap is a global method, which aims to preserve the geometry at all scales by mapping the nearby data points on the manifold to nearby points in the low-dimensional space and the faraway points to faraway points. In contrast with Isomap, LLE and LE are local methods, which assume that the global structure can be uncovered by keeping all the local structures of the dataset, and thus only attempt to preserve the local geometry.

As the manifold assumption matches the distribution of multimedia data very well, manifold learning has played an important role in various applications of multimedia data analysis. Pless explored the video representation via Isomap [32]. Rahimi et al. investigated the appearance manifolds from video by exploiting the temporal coherence between frames and supervision from a user [34]. Yang et al. learned the hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval [45]. Almeida et al. addressed the problem of video genre retrieval via reciprocal kNN graph based manifold learning [1, 31]. Shnitzer et al. applied a manifold learning technique, diffusion maps [5], to music analysis applications [38]. In our previous work, we proposed the hybrid manifold embedding for image classification [25]. Recently, we utilized supervised manifold regression to learn the compact representations of image/video data for media interestingness prediction [24].

2.2 Multi-view Dimensionality Reduction

Multi-view dimensionality reduction aims to find a common subspace for multiple views, where the cross-view correlation is preserved. The most classical method for multi-view dimensionality reduction is canonical correlation analysis (CCA), which aims to find the subspace where the correlations of data from two different views are maximized [15]. In [28], CCA was extended for multi-view scenarios. Luo et al. further proposed the tensor CCA for multi-view dimensionality reduction [27]. Zhang et al. presented a multi-view dimensionality co-reduction to explore the correlations within each view and maximize the dependence among different views [47].

Following the success of unsupervised multi-view dimensionality reduction, semi-supervised and supervised dimensionality reduction has been generalized for multi-view learning as well. Qian et al. introduced a semi-supervised multi-label multi-view dimensionality reduction framework based on reconstruction error minimization [33]. Wang et al. proposed a deep learning based multi-view dimensionality reduction scheme [42]. Kan et al. extended the classical linear discriminant analysis to the multi-view version [18].

Besides above general-purpose multi-view dimensionality reduction models, some methods have been presented for more specific applications in media data analysis. Li et al. studied the cross-view similarity search problem by mapping multiple views into a common Hamming space [21]. Farfade et al. modelled the problem of multi-view face detection using deep convolutional neural networks [10]. Jing et al. addressed the problem of web page classification using a semi-supervised intra-view and inter-view dimensionality reduction algorithm [17]. Zhu et al. proposed a block-row sparse multi-view multi-label learning method for image classification [48].

3 MULTI-VIEW MANIFOLD LEARNING

3.1 Problem Formulation

Let $\mathcal{X}$ be the set of data samples from V views: $\mathcal{X} = \{(\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1V}), (\mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2V}), ..., (\mathbf{x}_{n1}, \mathbf{x}_{n2}, ..., \mathbf{x}_{nV})\}$, where $\mathbf{x}_{iv} \in \mathbb{R}^{D_v}$ $(i = 1, ..., n,\ v = 1, ..., V)$ denotes the i-th sample from the v-th view, n denotes the number of data samples in the set, V denotes the number of views, and D_v denotes the original dimension of the v-th view. To represent the data from different views, we define V data matrices: $\mathbf{X}_1 = [\mathbf{x}_{11}, \mathbf{x}_{21}, ..., \mathbf{x}_{n1}]$, $\mathbf{X}_2 = [\mathbf{x}_{12}, \mathbf{x}_{22}, ..., \mathbf{x}_{n2}]$, ..., and $\mathbf{X}_V = [\mathbf{x}_{1V}, \mathbf{x}_{2V}, ..., \mathbf{x}_{nV}]$. Meanwhile, we have the corresponding label of each of the n data samples:

$\mathbf{l} = [l_1, l_2, ..., l_n]$, where $l_i \in [0, 1]$ is a real number denoting the corresponding label of the i-th data sample, 1 for extremely interesting and 0 for extremely non-interesting.

Given the above training set, Multi-view Manifold Learning (M^{2}L) aims to learn V transformation matrices $\mathbf{W}_v \in \mathbb{R}^{D_v \times d}$ ($d \ll D_v, v = 1, ..., V$), which are capable of projecting the original high-dimensional data from different views to a common subspace $\mathcal{Z} = \mathbb{R}^d$, where the manifold structure of the dataset could be well preserved, and the discriminant information could be maximized.

3.2 Geometry Preserving Criteria

In order to preserve the geometric structure of the dataset, we take both locality and globality into consideration. To describe the local structure for each view, we define a neighborhood graph matrix $\mathbf{N}^v$ for the v-th view ($v = 1, ..., V$). The element on the i-th row and the j-th column of $\mathbf{N}^v$ is given as follows:

$$N_{ij}^v = e^{-\frac{||\mathbf{x}_{iv} - \mathbf{x}_{jv}||^2}{2\sigma}}, \tag{1}$$

where the bandwidth σ is empirically set by $\sigma = \sum_{i=1}^{n} ||\mathbf{x}_{iv} - \mathbf{x}_{iv_K}||^2/n$ with $\mathbf{x}_{iv_K}$ being the K-th nearest neighbor of $\mathbf{x}_{iv}$. Accordingly, we construct the locality scatter matrix $\mathbf{S}_{lv}$ for the v-th view:

$$\mathbf{S}_{lv} = \sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}^v (\mathbf{x}_{iv} - \mathbf{x}_{jv})(\mathbf{x}_{iv} - \mathbf{x}_{jv})^T. \tag{2}$$

We then introduce the objective:

$$\{\mathbf{W}_1, ..., \mathbf{W}_V\} = \underset{\mathbf{W}_1,...,\mathbf{W}_V}{\arg\min}\ tr\Big(\sum_{v=1}^{V} \mathbf{W}_v^T \mathbf{S}_{lv} \mathbf{W}_v\Big), \tag{3}$$

where $tr(\cdot)$ denotes the trace operation. By minimizing the objective function in Eq. (3), the nearby data points in the original space of each view will also be close to each other in the common subspace, and thus the local structure is expected to be well captured.

To represent the global structure of each view, we construct the globality scatter matrix $\mathbf{S}_{gv}$:

$$\mathbf{S}_{gv} = \sum_{i=1}^{n}\sum_{j=1}^{n} (1 - N_{ij}^v)(\mathbf{x}_{iv} - \mathbf{x}_{jv})(\mathbf{x}_{iv} - \mathbf{x}_{jv})^T. \tag{4}$$

It is obvious that the larger the distance between $\mathbf{x}_{iv}$ and $\mathbf{x}_{jv}$, the smaller the N_{ij}^v, hence, the larger the $(1 - N_{ij}^v)$. This fact indicates that $\mathbf{S}_{gv}$ pays more attention to the relationship between faraway data points, which characterizes the structure of the v-th view from the global perspective. We then introduce the following objective:

$$\{\mathbf{W}_1, ..., \mathbf{W}_V\} = \underset{\mathbf{W}_1,...,\mathbf{W}_V}{\arg\max}\ tr\Big(\sum_{v=1}^{V} \mathbf{W}_v^T \mathbf{S}_{gv} \mathbf{W}_v\Big). \tag{5}$$

By maximizing the objective function in Eq. (5), the distant data points in the original space are still faraway from each other in the common subspace, and thus the global structure of the dataset is preserved.

3.3 Discrimination Maximizing Criteria

The idea behind the discriminant information maximization is straightforward: if two data points (no matter within the same view or from different views) possess similar interestingness levels, they should be close to each other in the learned subspace; otherwise, they should be faraway from each other. The more similar the interestingness levels, the closer the low-dimensional data points should be.

In order to describe the similarity between interestingness levels of different data samples, we define the label similarity matrix $\mathbf{A}$, where the element on the i-th row and the j-th column of $\mathbf{A}$ is given as follows:

$$A_{ij} = 1 - |l_i - l_j| \in [0, 1]. \tag{6}$$

Note that the representations of different views of the same data sample share the same label.

We then propose the following objective:

$$\begin{aligned}\{\mathbf{W}_1, ..., \mathbf{W}_V\} = \underset{\mathbf{W}_1,...,\mathbf{W}_V}{\arg\min}\ tr\Big(\sum_{u=1}^{V}\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}(\mathbf{W}_u^T \mathbf{x}_{iu} \\ - \mathbf{W}_v^T \mathbf{x}_{jv})(\mathbf{W}_u^T \mathbf{x}_{iu} - \mathbf{W}_v^T \mathbf{x}_{jv})^T\Big).\end{aligned} \tag{7}$$

By minimizing the objective function in Eq. (7), the data samples with similar or even the same interestingness levels will be close to each other in the learned common subspace, and thus the label similarity is well maintained.

To characterize the dissimilarity between data points, we define the discriminant matrix $\mathbf{B}$, where the element on the i-th row and the j-th column of $\mathbf{B}$ is given as follows:

$$B_{ij} = 1 - A_{ij} = |l_i - l_j| \in [0, 1]. \tag{8}$$

Then we present the following objective:

$$\begin{aligned}\{\mathbf{W}_1, ..., \mathbf{W}_V\} = \underset{\mathbf{W}_1,...,\mathbf{W}_V}{\arg\max}\ tr\Big(\sum_{u=1}^{V}\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}(\mathbf{W}_u^T \mathbf{x}_{iu} \\ - \mathbf{W}_v^T \mathbf{x}_{jv})(\mathbf{W}_u^T \mathbf{x}_{iu} - \mathbf{W}_v^T \mathbf{x}_{jv})^T\Big).\end{aligned} \tag{9}$$

By maximizing the objective function in Eq. (9), the data samples with different labels are expected to be faraway from each other in the learned common space, and thus the discriminant information can be well preserved.

3.4 Objective Function of M^{2}L

In order to preserve the geometric structure while maximizing the discriminant information, we integrate aforementioned four objectives (Eqs. (3), (5), (7), and (9)) into a unified formulation in Eq. (10), where $\alpha \in [0, 1]$ is a trade-off parameter to balance the weight of geometry information and that of discriminant information in the proposed learning framework.

$$\{\mathbf{W}_1, ..., \mathbf{W}_V\} = \underset{\substack{\mathbf{W}_1,...,\mathbf{W}_V \\ \mathbf{W}_v^T\mathbf{W}_v=\mathbf{I}_d \\ v=1,...,V}}{\arg\max}\ tr\left(\frac{\alpha\sum_{v=1}^{V}\mathbf{W}_v^T\mathbf{S}_{gv}\mathbf{W}_v + (1-\alpha)\sum_{u=1}^{V}\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j=1}^{n}B_{ij}(\mathbf{W}_u^T\mathbf{x}_{iu} - \mathbf{W}_v^T\mathbf{x}_{jv})(\mathbf{W}_u^T\mathbf{x}_{iu} - \mathbf{W}_v^T\mathbf{x}_{jv})^T}{\alpha\sum_{v=1}^{V}\mathbf{W}_v^T\mathbf{S}_{lv}\mathbf{W}_v + (1-\alpha)\sum_{u=1}^{V}\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}(\mathbf{W}_u^T\mathbf{x}_{iu} - \mathbf{W}_v^T\mathbf{x}_{jv})(\mathbf{W}_u^T\mathbf{x}_{iu} - \mathbf{W}_v^T\mathbf{x}_{jv})^T}\right) \tag{10}$$

3.5 Analytical Solution of M^{2}L

Actually, Eq. (10) could be rewritten as follows:

$$\mathbf{W} = \underset{\substack{\mathbf{W} \\ \mathbf{W}^T\mathbf{W}=\mathbf{I}_d}}{\arg\max}\ tr\left(\frac{\mathbf{W}^T\big(\alpha\mathbf{S}_g + (1-\alpha)\mathbf{Q}_B\big)\mathbf{W}}{\mathbf{W}^T\big(\alpha\mathbf{S}_l + (1-\alpha)\mathbf{Q}_A\big)\mathbf{W}}\right), \tag{11}$$

where $\mathbf{W} = [\mathbf{W}_1^T, \mathbf{W}_2^T, \dots, \mathbf{W}_V^T]^T$, $\mathbf{S}_g = diag(\mathbf{S}_{g1}, \mathbf{S}_{g2}, \dots, \mathbf{S}_{gV})$, $\mathbf{S}_l = diag(\mathbf{S}_{l1}, \mathbf{S}_{l2}, \dots, \mathbf{S}_{lV})$, $\mathbf{Q}_B = \begin{bmatrix} \mathbf{X}_1\mathbf{L}_{B11}\mathbf{X}_1^T & \cdots & \mathbf{X}_1\mathbf{L}_{B1V}\mathbf{X}_V^T \\ \vdots & \ddots & \vdots \\ \mathbf{X}_V\mathbf{L}_{BV1}\mathbf{X}_1^T & \cdots & \mathbf{X}_V\mathbf{L}_{BVV}\mathbf{X}_V^T \end{bmatrix}$, and $\mathbf{Q}_A = \begin{bmatrix} \mathbf{X}_1\mathbf{L}_{A11}\mathbf{X}_1^T & \cdots & \mathbf{X}_1\mathbf{L}_{A1V}\mathbf{X}_V^T \\ \vdots & \ddots & \vdots \\ \mathbf{X}_V\mathbf{L}_{AV1}\mathbf{X}_1^T & \cdots & \mathbf{X}_V\mathbf{L}_{AVV}\mathbf{X}_V^T \end{bmatrix}$. Here $\mathbf{L}_{Buv}$ and $\mathbf{L}_{Auv}$ $(u, v = 1, ..., V)$ are the (u, v)-th blocks of $\mathbf{L}_B$ and $\mathbf{L}_A$, respectively, i.e., $\mathbf{L}_B = \begin{bmatrix} \mathbf{L}_{B11} & \cdots & \mathbf{L}_{B1V} \\ \vdots & \ddots & \vdots \\ \mathbf{L}_{BV1} & \cdots & \mathbf{L}_{BVV} \end{bmatrix}$ and $\mathbf{L}_A = \begin{bmatrix} \mathbf{L}_{A11} & \cdots & \mathbf{L}_{A1V} \\ \vdots & \ddots & \vdots \\ \mathbf{L}_{AV1} & \cdots & \mathbf{L}_{AVV} \end{bmatrix}$, where $\mathbf{L}_B$ and $\mathbf{L}_A$ are two $nV \times nV$ Laplacian matrices [2] defined as $\mathbf{L}_B = \mathbf{Diag}_B - \mathbf{Adj}_B$ and $\mathbf{L}_A = \mathbf{Diag}_A - \mathbf{Adj}_A$, in which $\mathbf{Adj}_B = \mathbf{1}_{V\times V} \otimes \mathbf{B}$ and $\mathbf{Adj}_A = \mathbf{1}_{V\times V} \otimes \mathbf{A}$ are also $nV \times nV$ matrices, $\mathbf{1}_{V\times V}$ denotes the $V \times V$ matrix with all entries being 1, the symbol $\otimes$ denotes the Kronecker product operation on two matrices, and $\mathbf{Diag}_B$ and $\mathbf{Diag}_A$ are diagonal matrices defined as $(Diag_B)_{ii} = \sum_{j=1}^{nV}(Adj_B)_{ij}$ and $(Diag_A)_{ii} = \sum_{j=1}^{nV}(Adj_A)_{ij}$, respectively $(i = 1, ..., nV)$. Then the optimal $\mathbf{W}$ that maximizes the objective function in Eq. (11) is composed of the eigenvectors corresponding to the d largest eigenvalues of the following generalized eigen-decomposition problem:

$$\big(\alpha\mathbf{S}_g + (1-\alpha)\mathbf{Q}_B\big)\mathbf{w} = \lambda\big(\alpha\mathbf{S}_l + (1-\alpha)\mathbf{Q}_A\big)\mathbf{w}. \tag{12}$$

The detailed procedure of M^{2}L is described in Algorithm 1.

Algorithm 1: Multi-view Manifold Learning (M^{2}L)

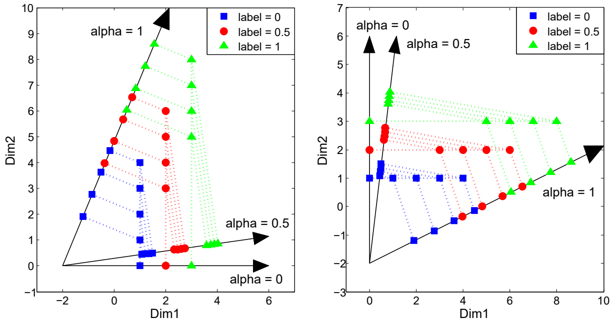
```
Input: The training dataset
    $\mathcal{X} = \{(\mathbf{x}_{11}, ..., \mathbf{x}_{1V}), ..., (\mathbf{x}_{n1}, ..., \mathbf{x}_{nV})\}$ and the
    corresponding label set $\mathcal{L} = \{l_1, ..., l_n\}$; the
    subspace dimension $d$; and the balancing
    parameter $\alpha$
Output: Transformation matrices $\mathbf{W}_1, \cdots, \mathbf{W}_V$
for $v = 1, ..., V$ do
    Construct $\mathbf{N}^v$ according to Eq. (1);
    Construct $\mathbf{S}_{lv}$ according to Eq. (2);
    Construct $\mathbf{S}_{gv}$ according to Eq. (4);
end
Construct $\mathbf{S}_l$: $\mathbf{S}_l \leftarrow diag(\mathbf{S}_{l1}, \mathbf{S}_{l2}, \dots, \mathbf{S}_{lV})$;
Construct $\mathbf{S}_g$: $\mathbf{S}_g \leftarrow diag(\mathbf{S}_{g1}, \mathbf{S}_{g2}, \dots, \mathbf{S}_{gV})$;
Construct $\mathbf{A}$ according to Eq. (6);
Construct $\mathbf{Adj}_A$: $\mathbf{Adj}_A \leftarrow \mathbf{1}_{V\times V} \otimes \mathbf{A}$;
Construct $\mathbf{Diag}_A$: $(Diag_A)_{ii} \leftarrow \sum_{j=1}^{nV}(Adj_A)_{ij}$;
Construct $\mathbf{L}_A$: $\mathbf{L}_A \leftarrow \mathbf{Diag}_A - \mathbf{Adj}_A$;
Construct $\mathbf{Q}_A$:
    $\mathbf{Q}_A = \begin{bmatrix} \mathbf{X}_1\mathbf{L}_{A11}\mathbf{X}_1^T & \cdots & \mathbf{X}_1\mathbf{L}_{A1V}\mathbf{X}_V^T \\ \vdots & \ddots & \vdots \\ \mathbf{X}_V\mathbf{L}_{AV1}\mathbf{X}_1^T & \cdots & \mathbf{X}_V\mathbf{L}_{AVV}\mathbf{X}_V^T \end{bmatrix}$;
Construct $\mathbf{B}$ according to Eq. (8);
Construct $\mathbf{Adj}_B$: $\mathbf{Adj}_B \leftarrow \mathbf{1}_{V\times V} \otimes \mathbf{B}$;
Construct $\mathbf{Diag}_B$: $(Diag_B)_{ii} = \sum_{j=1}^{nV}(Adj_B)_{ij}$;
Construct $\mathbf{L}_B$: $\mathbf{L}_B \leftarrow \mathbf{Diag}_B - \mathbf{Adj}_B$;
Construct $\mathbf{Q}_B$:
    $\mathbf{Q}_B = \begin{bmatrix} \mathbf{X}_1\mathbf{L}_{B11}\mathbf{X}_1^T & \cdots & \mathbf{X}_1\mathbf{L}_{B1V}\mathbf{X}_V^T \\ \vdots & \ddots & \vdots \\ \mathbf{X}_V\mathbf{L}_{BV1}\mathbf{X}_1^T & \cdots & \mathbf{X}_V\mathbf{L}_{BVV}\mathbf{X}_V^T \end{bmatrix}$;
Obtain $\mathbf{W}$ by solving the generalized eigen-
decomposition problem:
    $\big(\alpha\mathbf{S}_g + (1-\alpha)\mathbf{Q}_B\big)\mathbf{w} = \lambda\big(\alpha\mathbf{S}_l + (1-\alpha)\mathbf{Q}_A\big)\mathbf{w}$;
$[\mathbf{W}_1^T, \mathbf{W}_2^T, \dots, \mathbf{W}_M^T] \leftarrow \mathbf{W}^T$;
```

4 EXPERIMENTAL RESULTS

4.1 Synthetic Example

First, we schematically illustrate the effect of M^{2}L on a synthetic dataset. This synthetic dataset is composed of twelve two-dimensional data samples with two views, where $\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 3 & 4 & 5 & 6 & 5 & 6 & 7 & 8 \end{bmatrix}$ and $\mathbf{X}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\mathbf{X}_1$, i.e., we interchange the two dimensions in view 1 to generate view 2. Moreover, we set the label vector $\mathbf{l} = [0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 1]$.

Figure 1 shows the projection directions learned by M^{2}L and the corresponding one-dimensional mapping results for these two views (Figure 1(a) for view 1 and Figure 1(b) for view 2). For each view, we demonstrate three projection



(a) (b)

Figure 1: Schematic illustration of M^{2}L on the synthetic example. (a) The original data in view 1 and the corresponding projections generated by M^{2}L. (b) The original data in view 2 and the corresponding projections generated by M^{2}L.

directions with $\alpha = 0$, 0.5, and 1 respectively. When $\alpha = 0$, the proposed M^{2}L considers only the discriminant information but ignores the geometric structure. Therefore, the data with different label values are well separated but the intra-class manifold information is lost in the one-dimensional subspace ($\mathbf{w}_1 = [1, 0]^T$ for view 1 and $\mathbf{w}_2 = [0, 1]^T$ for view 2). In contrast, when $\alpha = 1$, i.e., ignoring the discriminant information, the mapping results preserve the manifold structure well but the data with different labels are overlapped ($\mathbf{w}_1 = [0.4146, 1]^T$ and $\mathbf{w}_2 = [1, 0.4146]^T$). By setting $\alpha = 0.5$, both geometric structure and discriminant information are taken into consideration, and thus are well preserved ($\mathbf{w}_1 = [1, 0.1441]^T$ and $\mathbf{w}_2 = [0.144, 1]^T$). Interestingly, the relation between two views (i.e., the row exchange) is faithfully reflected by the corresponding projection vectors. Moreover, in the one-dimensional subspace, the mapping values of data from different views are similar if they possess the same label value; and are distinct from each other if the label values are different. This indicates that the learned common subspace preserves the label information consistently across different views.

4.2 Media Interestingness Prediction

In this subsection, we evaluate the performance of the proposed method on the Predicting Media Interestingness dataset [8]. The data is extracted from 78 Creative Commons licensed trailers of Hollywood-like movies. The trailers are further segmented into 7,396 shots, 5,054 of which are used for training and the remaining 2,342 shots are used for test. For each shot, the middle frame is extracted as its key frame and a set of low-level features of this key frame is computed as the original feature representation of the corresponding shot [8].

In our experiments, we utilize five types of features, i.e., 300-D dense Scale Invariant Feature Transform (dense SIFT) features [20, 26], 300-D HoG 2×2 features [7, 43], 59-D Local Binary Patterns (LBP) features [29], 512-D GIST features [30], and 128-D color histogram features. We thus have 5 views and the total dimension of these 5 views is 1,299. All data are manually annotated in terms of interestingness ranging from 0 to 1 by human assessors.

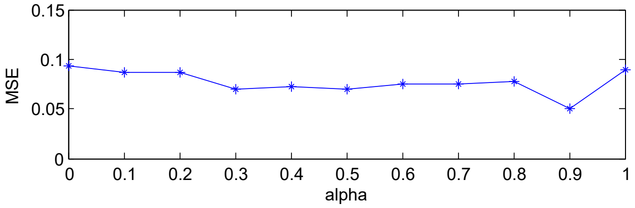


Figure 2: MSE of M^{2}L under various values of α.

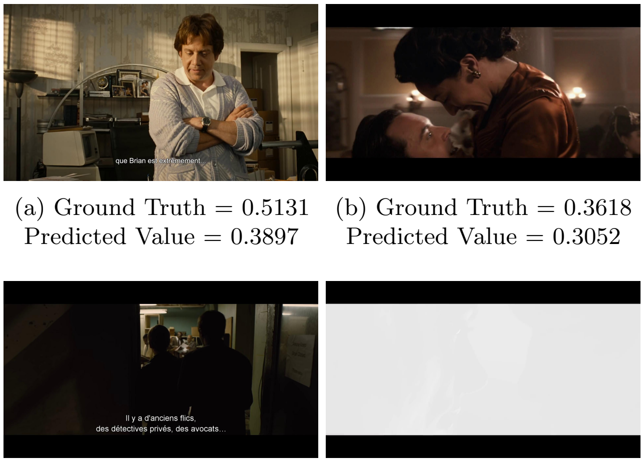
Table 1: Performance (in terms of MSE) of PCA, MCCA, SMR, and M^{2}L with the corresponding reduced dimensions on Predicting Media Interestingness dataset

	PCA	MCCA	SMR	M^{2}L
MSE	0.1850	0.1197	0.0599	**0.0340**
Red. Dim.	37	$3 \times 5 = 15$	15	$\mathbf{2 \times 5 = 10}$

Firstly, we compare the proposed method M^{2}L with three dimensionality reduction methods, i.e., principal component analysis (PCA) [14], multi-view CCA (MCCA) [28], and supervised manifold regression (SMR) [24], where PCA is the most classical dimensionality reduction method, MCCA is a representative multi-view dimensionality reduction algorithm, and SMR is a recently proposed dimensionality reduction method for continuous labels. For MCCA and M^{2}L, we first map the original representations of 5 views to a low-dimensional common subspace, and then concatenate the reduced representations of these 5 views to a single feature vector. For PCA and SMR, we first concatenate the original representations of these 5 views to a 1,299-dimensional feature vector, and then map it to a low-dimensional subspace. After dimensionality reduction, we utilize ϵ-SVR on the low-dimensional data to predict the interestingness levels. We choose LIBSVM [4] and the RBF kernel, and the parameters in ϵ-SVR for the low-dimensional representations generated by PCA, MCCA, SMR, and M^{2}L are determined separately by a 5-fold cross validation technique on training data.

We also use the 5-fold cross validation technique on training data to determine the optimal value of α for M^{2}L. From Figure 2 we can see that the mean square error (MSE) value is relatively large when $\alpha = 0$ (preserving only the discriminant information) or $\alpha = 1$ (preserving only the geometric structure). The performance becomes better when both the geometric structure and the discriminant information are taken into consideration, and achieves the best when $\alpha = 0.9$. Therefore, we set $\alpha = 0.9$ for M^{2}L in the remaining experiments.

Table 1 reports the best performance (in terms of MSE, i.e., mean squared error) and the corresponding reduced dimension of aforementioned four algorithms. By learning


que Brian est extrêmement
Il y a d'anciens flics,
des détectives privés, des avocats...

(a) Ground Truth = 0.5131
Predicted Value = 0.3897

(b) Ground Truth = 0.3618
Predicted Value = 0.3052

(c) Ground Truth = 0.0749
Predicted Value = 0.0735

(d) Ground Truth = 0.0559
Predicted Value = 0.0575

Figure 3: Examples of interesting and non-interesting video frames from the Predicting Media Interestingness dataset. (a) and (b) are examples of video frames with high interestingness values while (c) and (d) are examples of video frames with low interestingness values.

Table 2: Performance (in terms of MSE) of M^2L on individual views of Predicting Media Interestingness dataset

	SIFT	HoG	LBP	GIST	ColorHist
MSE	0.0462	0.0522	0.0813	0.0723	0.0360
Red. Dim.	2	2	2	2	2

the intrinsic features from multiple views jointly, MCCA and M^2L outperform the classical PCA. By preserving the geometric structure and the discriminant information of the dataset, SMR and M^2L achieve better results than PCA and MCCA. By integrating the geometric structures of different views and the interestingness information of the original dataset under a multi-view manifold learning framework, M^2L performs the best in this experiment. Moreover, the reduced dimension of M^2L is the lowest among four methods, which indicates that the proposed method is capable of capturing the interestingness information of original dataset in a very low-dimensional subspace.

Figure 3 illustrates some examples of interesting and non-interesting video frames from the Predicting Media Interestingness dataset. Images in the top row of Figure 3 are two examples of video frames with high interestingness values and images in the bottom row are two examples of video frames with low interestingness values. We also list the ground truth and the predicted value on these video frames in the figure. We can observe that our predictions on these frames are consistent with the human judgment in terms of interestingness.

After having demonstrated the overall performance of M^2L on all five views, we further measure the MSE values of M^2L on individual views. Table 2 gives the results of M^2L on dense SIFT, HoG 2×2, LBP, GIST, and color histogram, respectively. We can observe that even using the low-dimensional representation from only one view can generate good result in predicting the media interestingness. Specifically, the performance of M^2L on LBP and GIST is better than that of PCA and MCCA with all five views; and the performance of M^2L on dense SIFT, HoG 2×2, and color histogram further outperforms SMR with all five views. The reason of achieving good performance on individual view might be that although only one view is used in prediction, the transformation matrix for that individual view is learned by considering the information from all the five views.

5 CONCLUSIONS

In this paper, we propose a novel Multi-view Manifold Learning (M^2L) algorithm to uncover the mapping between visual features and the corresponding interestingness levels. By modelling both geometric structures and interestingness information in a multi-view learning framework, the proposed algorithm shows good and explainable results in both synthetic example and real-world dataset.

For the future work, we are interested in mining the physical meaning of each dimension in the learned subspace, as well as its relations with the findings from psychology. Moreover, we aim to improve the proposed model by considering the high-order tensor structure and the dynamic nature of video data. Last but not least, we are going to investigate the possibility of extending the current method to the imbalanced version as most of the users only pay attention to the interesting media data rather than the non-interesting ones.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments and suggestions. This work was supported in part by the National Natural Science Foundation of China under Grants 61503317 and 61672444, in part by the Faculty Research Grant of Hong Kong Baptist University (HKBU) under Project FRG2/16-17/032, in part by the Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation (No. JTKF2017001), Changsha University of Science & Technology, P.R. China, and in part by the Science and Technology Research and Development Fund of Shenzhen with Project Code JCYJ20160531194006833.

REFERENCES

[1] J. Almeida, D. C. G. Pedronette, and O. A. B. Penatti. 2014. Unsupervised manifold learning for video genre retrieval. In *Proc. CIARP*. 604–612.

[2] M. Belkin and P. Niyogi. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* 15, 6 (2003), 1373–1396.

[3] P. R. Chakraborty, D. Tjondronegoro, L. Zhang, and V. Chandran. 2016. Automatic Identification of Sports Video Highlights Using Viewer Interest Features. In *Proc. ACM ICMR*. 55–62.

[4] C.-C. Chang and C.-J. Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2, 3 (2011), 27:1–27:27.

[5] R. R. Coifman and S. Lafon. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis* 21, 1 (2006), 5 – 30.

[6] M. G. Constantin, B. Boteanu, and B. Ionescu. 2016. LAPI at MediaEval 2016 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2016 Workshop*.

[7] N. Dalal and B. Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proc. CVPR*. 886–893.

[8] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre. 2016. MediaEval 2016 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2016 Workshop*.

[9] S. Dhar, V. Ordonez, and T. L. Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*. 1657–1664.

[10] S. S. Farfade, M. J. Saberian, and L.-J. Li. 2015. Multi-view Face Detection Using Deep Convolutional Neural Networks. In *Proc. 5th ACM ICMR*. 643–650.

[11] L. Geng and H. J. Hamilton. 2006. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.* 38, 3 (2006).

[12] H. Grabner, F. Nater, M. Druey, and L. Van Gool. 2013. Visual Interestingness in Image Sequences. In *Proc. 21st ACM Multimedia*. 1017–1026.

[13] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. 2013. The Interestingness of Images. In *Proc. ICCV*. 1633–1640.

[14] H. Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 6 (1933), 417–441.

[15] H. Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28 (1936), 321–377.

[16] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. 2013. Understanding and Predicting Interestingness of Videos. In *Proc. 27th AAAI*.

[17] X.-Y. Jing, Q. Liu, F. Wu, B. Xu, Y. Zhu, and S. Chen. 2015. Web Page Classification Based on Uncorrelated Semi-supervised Intra-view and Inter-view Manifold Discriminant Feature Extraction. In *Proc. 24th IJCAI*. 2255–2261.

[18] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. 2016. Multi-View Discriminant Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1 (2016), 188–194.

[19] H. Katti, K. Y. Bin, T.-S. Chua, and M. Kankanhalli. 2008. Pre-attentive discrimination of interestingness in images. In *Proc. ICME*. 1433–1436.

[20] S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. CVPR*. 2169–2178.

[21] K. Li, G. Qi, J. Ye, and K. Hua. 2016. Linear Subspace Ranking Hashing for Cross-modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* PP, 99 (2016), 1–1.

[22] C. Liem. 2016. TUD-MMC at MediaEval 2016: Predicting Media Interestingness Task. In *Proc. of the MediaEval 2016 Workshop*.

[23] F. Liu, Y. Niu, and M. Gleicher. 2009. Using Web Images for Measuring Video Frame Interestingness. In *Proc. 21st IJCAI*.

[24] Y. Liu, Z. Gu, and Y.-M. Cheung. 2016. Supervised Manifold Learning for Media Interestingness Prediction. In *Proc. of the MediaEval 2016 Workshop*.

[25] Y. Liu, Y. Liu, K. C. C. Chan, and K. A. Hua. 2014. Hybrid Manifold Embedding. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 12 (2014), 2295–2302.

[26] D. G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[27] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen. 2015. Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction. *IEEE Trans. Knowl. Data Eng.* 27, 11 (2015), 3111–3124.

[28] A.A. Nielsen. 2002. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans. Image Process.* 11, 3 (2002), 293–305.

[29] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (2002), 971–987.

[30] A. Oliva and A. Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vision* 42, 3 (2001), 145–175.

[31] D. C. G. Pedronette, O. A.B. Penatti, and R. da S. Torres. 2014. Unsupervised manifold learning using Reciprocal kNN Graphs in image re-ranking and rank aggregation tasks. *Image and Vision Computing* 32, 2 (2014), 120–130.

[32] R. Pless. 2003. Image spaces and video trajectories: using Isomap to explore video sequences. In *Proc. ICCV*. 1433–1440.

[33] B. Qian, X. Wang, J. Ye, and I. Davidson. 2015. A Reconstruction Error Based Framework for Multi-Label and Multi-View Learning. *IEEE Trans. Knowl. Data Eng.* 27, 3 (2015), 594–607.

[34] A. Rahimi, B. Recht, and T. Darrell. 2005. Learning Appearance Manifolds from Video. In *Proc. CVPR*. 868–875.

[35] S. Rayatdoost and M. Soleymani. 2016. Ranking Images and Videos on Visual Interestingness by Visual Sentiment Features. In *Proc. of the MediaEval 2016 Workshop*.

[36] S. Roweis and L. K. Saul. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290, 5500 (2000), 2323–2326.

[37] Y. Shen, C.-H. Demarty, and N. Q. K. Duong. 2016. Technicolor at MediaEval 2016 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2016 Workshop*.

[38] T. Shnitzer, R. Talmon, and J. J. Slotine. 2017. Manifold Learning With Contracting Observers for Data-Driven Time-Series Analysis. *IEEE Trans. Signal Process.* 65, 4 (2017), 904–918.

[39] M. Soleymani. 2015. The Quest for Visual Interest. In *Proc. 23rd ACM Multimedia*. 919–922.

[40] J. B. Tenenbaum, V. de Silva, and J. C. Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 5500 (2000), 2319–2323.

[41] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. V. Gool. 2016. ETH-CVL at MediaEval 2016: Textual-Visual Embeddings and Video2GIF for Video Interestingness. In *Proc. of the MediaEval 2016 Workshop*.

[42] W. Wang, R. Arora, K. Livescu, and J. Bilmes. 2015. On Deep Multi-View Representation Learning. In *Proc. 32nd ICML*. 1083–1092.

[43] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*. 3485–3492.

[44] B. Xu, Y. Fu, and Y.-G. Jiang. 2016. BigVid at MediaEval 2016: Predicting Interestingness in Images and Videos. In *Proc. of the MediaEval 2016 Workshop*.

[45] Y. Yang, Y. T. Zhuang, F. Wu, and Y. H. Pan. 2008. Harmonizing Hierarchical Manifolds for Multimedia Document Semantics Understanding and Cross-Media Retrieval. *IEEE Trans. Multimedia* 10, 3 (2008), 437–446.

[46] G. Zen, P. de Juan, Y. Song, and A. Jaimes. 2016. Mouse Activity As an Indicator of Interestingness in Video. In *Proc. ACM ICMR*. 47–54.

[47] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. 2017. Flexible Multi-View Dimensionality Co-Reduction. *IEEE Trans. Image Process.* 26, 2 (2017), 648–659.

[48] X. Zhu, X. Li, and S. Zhang. 2016. Block-Row Sparse Multiview Multilabel Learning for Image Classification. *IEEE Trans. Cybern.* 46, 2 (2016), 450–461.