Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: http://ees.elsevier.com

MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging

Mona Zamiri [a,b,1], Tahereh Bahraini [b,c,2], Hadi Sadoghi Yazdi [a,b,3,*]

[a] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
[b] Center of Excellence on Soft Computing and Intelligent Information Processing, Ferdowsi University of Mashhad, Mashhad, Iran
[c] Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Article history:
Received 23 November 2020
Received in revised form 14 January 2021
Accepted 23 January 2021
Available online xxx

Keywords
Multi-view spectral clustering
Image annotation
Geo-tagged photos
Image tagging
Recommender systems
Geographical information

ABSTRACT

Image tag recommendation, aiming at assigning a set of relevant tags for images, is a useful way to help users organize images' content. Early methods in image tagging mainly demonstrated using low-level visual features. However, two visually similar photos may have different concepts (semantic gap). Although different multi-view tagging methods are proposed to learn the discriminative features, they usually do not consider the geographical correlation among images. Moreover, geographical-based image tagging models generally focused on the relevance criterion, i.e., how well the suggested tags describe image content. Diversity and redundancy should be controlled to guarantee the recommendation models' effectiveness and promote complementary information among tags. This paper proposes a robust multi-view image tagging method, termed MVDF-RSC, which considers the relevance, diversity, and redundancy criteria. Precisely, the proposed method consists of two phases: training and prediction. We propose a new robust optimization problem in the training phase to determine the similarity between data via the early fusion of multiple views of images and obtain clusters. In the prediction phase, relevant tags are recommended to each test data using a search-based method and a late fusion strategy. Comprehensive experiments on two geo-tagged image datasets demonstrate the proposed method's effectiveness over state-of-the-art alternatives.

1. Introduction

Nowadays, digital cameras and Internet technologies have led to the noticeable growth of applications and platforms for users to communicate with their peers and produce various contents- e.g., images or video clips. Social contents generally have rich metadata like geospatial information, tags, and visual features. Thanks to the improvement of GPS-enabled devices and digital imaging technologies, among different kinds of metadata going through social media websites, geographical information is of great interest. Geo-tagged contents provide an opportunity for researchers to propose different applications ranging from traffic analysis (Jia, Khadka, & Kimc, 2018; Wang, Li, & Zhu, 2020) and itinerary recommender systems (Kou, Leong Hou, Yang, & Gong, 2015; Cai, Lee, & Lee, 2018; Jiang, Yin, Wang, & Yu, 2013) to the point of interest (POI) discovery (Qian et al., 2020; Kuo, Chan, Fan, & Zipf, 2018; Xing, Meng, Hou, Song, & Xu, 2017). Also, spatial analysis has been employed in healthcare research (Boulos, Peng, & VoPham, 2019; Soltanisehat, Alizadeh, Hao, & Choo, 2020).

Recommender systems have been proposed to organize and manage various multimedia content (Deldjoo, Schedl, Cremonesi, & Pasi, 2020; Logesh, Subramaniyaswamy, Malathi, Sivaramakrishnan, & Vijayakumar, 2020). They attempt to recommend items of different media types, such as text and image. Due to the dramatic growth of digital photos, image tag recommendation methods are proposed to assign relevant tags to the images and help users organize and index photos' content (Lei, Liu, & Li, 2016; Nwana & Chen, 2017; Li, Shi, Du, Liu, & Wen, 2016). Early image tagging methods provided a model that utilizes low-level visual features, such as color, shape, and texture, in order to determine suitable tags. However, there is a challenge, semantic gap between low-level visual features and high-level concepts. For instance, these models cannot distinguish between "sky" and "water" due to the similar texture and color features. Multi-view data fusion techniques have attracted vast attention in image tagging over recent decades to bridge the semantic gap (Rad & Jamzad, 2017; Rad & Jamzad, 2015; Kalayeh, Idrees, & Shah, 2014; Xue, Li, & Huang, 2018). Multi-view data fusion can be defined as combining various views to accomplish different multimedia data mining tasks

* Corresponding author at: Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
E-mail addresses: zamiri.mona@mail.um.ac.ir (M. Zamiri); bahraini.tahereh@mail.um.ac.ir (T. Bahraini); h-sadoghi@um.ac.ir (H.S. Yazdi)
[1] ORCID ID: https://orcid.org/0000-0001-8315-7027.
[2] ORCID ID: https://orcid.org/0000-0003-0130-1764.
[3] ORCID ID: https://orcid.org/0000-0002-6885-4956.

https://doi.org/10.1016/j.eswa.2021.114657
0957-4174/

(Atrey, Hossain, Saddik, & Kankanhalli, 2010). These models generally focus on multiple low-level visual features and learning textual feature space and visual feature space without considering geographical information, in which a few tags can be well recommended in real-world applications. There are essential tags in social datasets, also called indirectly content-related tags (Lei et al., 2016), which indirectly describe images' content. For example, the tag "Mexico" in Fig. 1 represents an image taken in Mexico City, while it is difficult to understand such tags from the low-level visual features of the images. It is believed that geospatial information (i.e., latitude and longitude) of photos is tied to the images' contents intimately (Lee, Won, & McLeod, 2008; Toyama, Logan, & Roseway, 2003). Knowing that an image was taken in Disneyland, for example, gives a lot of information about the image before viewing a single pixel of it. Most of the previous image tagging methods assumed that images are independently and identically distributed, and it is not useful to capture the geospatial information correlation. However, this approach is insufficient for real-world data where there are indirectly content-related tags. Different image tagging methods have been proposed that employ the correspondence between visual features of each image and its geospatial information (Zheng, Caiming, & Caixian, 2018; Zhang, Zhao, Zhang, Wang, & Li, 2018; Qian, Liu, Zheng, Du, & Hou, 2013; Liu, Li, Tang, Jiang, & Lu, 2014; Zamiri & Yazdi, 2020). However, these geo-based image tagging models generally focused on the relevance criterion of the recommended tags. In image tagging, this criterion refers to how well the suggested tags describe the target image's content. Nevertheless, the relevance criterion only may not guarantee the effectiveness of recommendation models. For example, consider a list of candidate tags to a specific image, in which all candidates may describe the content of that image, but all of them are synonyms or redundant. In multimedia, where images may be multifaceted, providing a tag recommendation model that controls diversity and redundancy, as well as relevance criterion, may promote more complementary information across different views, helping to cover various aspects related to target image and, finally, enhancing the performance of tagging method. Also, most of the prior geo-based image tagging methods are not sufficient for multimedia datasets generated by users and may contain noise. Although we proposed a robust multi-view model for image tagging in (Zamiri & Yazdi, 2020), there is still much room for improvement. First, this graph-based model considers only the relevance criterion. Second, it conducts graph learning and clustering step separately. It does not consider the graph's quality since it does not pay attention to the fact that the fusion graph should have exact k – the number of clusters – connected components. Thus, this graph might lead to suboptimal clustering results.

Fig. 1. An example image from Flickr website with its corresponding tags.

This paper puts forward a graph-based multi-view model for the image tagging problem by learning a robust fusion similarity graph of different views. The proposed method considers relevance, diversity, and redundancy for fusion graph learning and simultaneously performs graph fusion and spectral clustering. Fig. 2 shows a schematic illustration of the proposed method. The contributions of this paper can be summarized as follows:

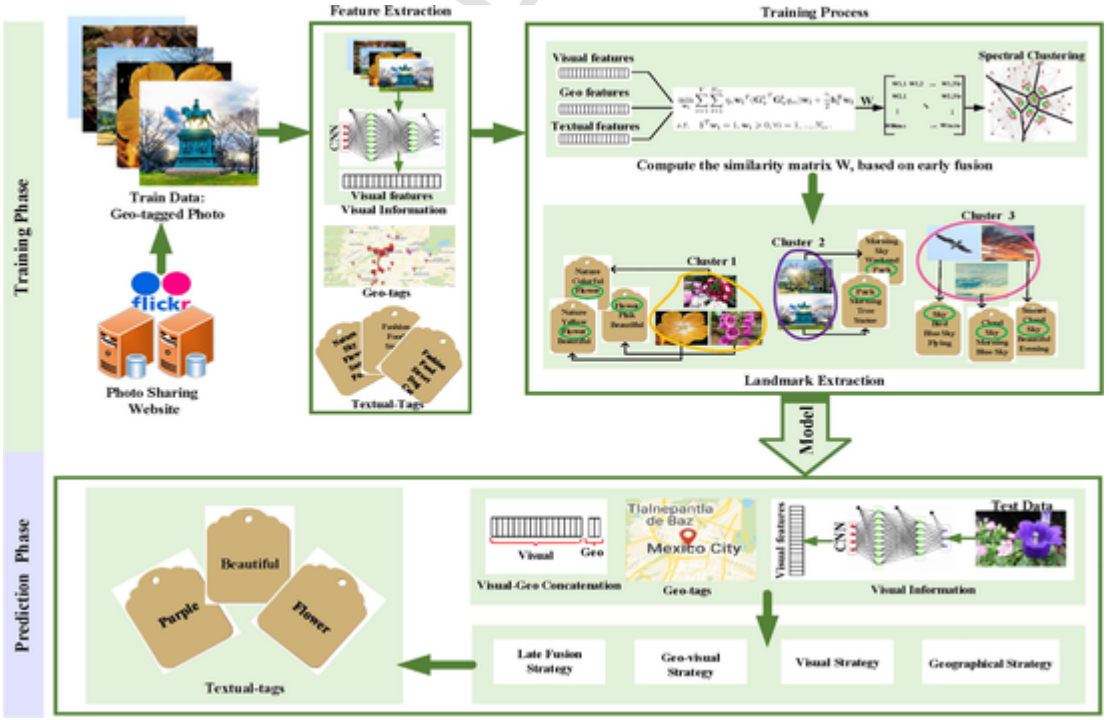

Fig. 2. Schematic illustration of the proposed image tagging method.

- We propose a novel multi-view robust spectral clustering method based on the maximum correntropy criterion (MCC) to determine the relationships between the training images and their corresponding concepts. This clustering method takes a new similarity measurement to integrate visual, textual, and geographical views of images.
- We use a diversity regularization term to learn the weights of various views adaptively. This regularization term is useful for enhancing diversity and suppressing the redundancy of different views.
- Our model finds clusters in the fusion step with no additional clustering step to promote the similarity graph learning.
- An effective fusion technique with early and late fusion is provided in the training and prediction phases respectively.
- Geographical information is employed to enhance the performance of the tagging model.

The rest of this paper is organized as follows. In section 2, we review the related work on image tagging, multi-view clustering techniques, and maximum correntropy criterion. Section 3 represents the proposed method. Experimental results and evaluations are shown in section 4. Conclusions are given in section 5.

2. Related work

In this section, we review the existing image tagging models, multi-view clustering methods, and maximum correntropy criterion.

2.1. Image tagging methods

In general, image tagging methods can be classified into three categories: (1) generative methods, (2) discriminative methods, and (3) search-based models. The first group tries to maximize the likelihood of features and tags and estimate the distribution parameters from training datasets. The generative tagging methods can be divided into three types, relevance model, topic model, and hidden Markov model. The relevance models generate a joint distribution over image features and tags and then try to compute each tag's posterior probability for test images. It then chooses a tag with the highest probability of a new image. Various relevance methods have been proposed for tagging tasks (Valcarce, Parapar, & Barreiro, 2018; Parapar, Bellogin, Castells, & Barreiro, 2013), including multiple Bernoulli relevance models (MBRM) (Feng, Manmatha, & Lavrenko, 2004) and Cross-Media Relevance Models (CMRM) (Jeon, Lavrenko, & Manmatha, 2003). The topic methods explore topics from tagged images, probability distributions over annotation tags and image features. Topic models typically consist of the latent semantic analysis (LSA) (Zhang et al., 2018), latent Dirichlet allocation (LDA), and probabilistic latent semantic analysis (pLSA). Approaches, such as LDA and pLSA, utilize the topic concept to handle the joint modeling of content and visual information(Zheng et al., 2018; Zhang, Zhang, Wang, & Guan, 2011; Putthividhy, Attias, & Nagarajan, 2010). Non-negative matrix factorization models fall into this group, like (Rad & Jamzad, 2015; Rad & Jamzad, 2017). The Hidden Markov model (HMM) is much popular in the tagging field (Yu & Ip, 2006; Ghoshal, Ircing, & Khudanpur, 2005; Zhao, Zhao, & Zhu, 2009).

The second group learns a classifier for each tag and determines the class of each input image. This classifier is learned based on different learning methods, for instance, Neural Networks (NN) (Hu, tong Zhou, Deng, Liao, & Mori, 2016; Savita, Patel, & Sinhal, 2013; Wang, Xie, Xue, & Zhang, 2017) or Support Vector Machine (SVM) (Verma & Jawahar, 2013). In recent years, many graph-based strategies that model data as a graph are provided (Amiri & Jamzad, 2018; Zhao, Chow, Zhang, & Li, 2015; Tian, Wang, Li, & Sun, 2019).

Like k-nearest neighbor, the search-based models retrieve relevant tags for each input query sample based on the tags of similar data (Makadia, Pavlovic, & Kumar, 2008). Some papers provide a distance metric learning (DML) scheme to compute the distance of features (Bar-Hillel, Hertz, Shental, & Weinshall, 2005; Weinberger & Saul, 2009; Xing, Ng, Jordan, & Russell, 2003).

Many papers have proposed a hybrid model that utilizes the advantages of more than one group (Altan & Karasu, 2020). Kalayeh et al. proposed the NMF-KNN method to solve a continuous increase in data and tags. It has been specifically learned a model for each image (Kalayeh et al., 2014). Image annotation using a generative model and a search-based algorithm is considered by (Rad & Jamzad, 2015). This method extracts the latent factors and represents data to low-rank latent factors space using NMF. It predicts tags using a search-based method in this space. They also proposed an annotation method, a hybrid model of generative and nearest neighbor-based methods (Rad & Jamzad, 2017). Their model finds a latent space by NMF and allows it to choose its number of basis factors for each view. Finally, a weighted nearest neighbor based on a unified distance matrix is applied to predict their tags for the query images. Li et al. provided a hybrid method that utilizes a probabilistic latent semantic analysis (PLSA) in the generative learning phase and an ensemble of classifiers to classify multi-label data in the discriminative learning stage (Li, Shi, Zhao, Li, & Tang, 2013). Murthy et al. introduced a hybrid method combining generative and discriminative models for image annotation (Murthy, Can, & Manmatha, 2014). Their method uses SVM to address the problem of irrelevant keywords. Also, a discrete multiple Bernoulli relevance model (DMBRM) is used to address imbalanced data.

Similar to (Zamiri & Yazdi, 2020), we propose a hybrid model that uses the discriminative and search-based models for image tagging. Nowadays, optimization problems play a pivotal role in different applications, like (Karasu, Altan, Bekiros, & Ahmad, 2020). We employ multiple views (i.e., visual, textual, and geographical features) in various matrices and propose an optimization problem based on the MCC to find the optimal fusion graph and clusters simultaneously. We add a diversity regularization term to the optimization problem of (Zamiri & Yazdi, 2020) and a rank constraint to reach an ideal fusion similarity graph. Then, we determine the landmarks, the most repetitive tags, for each cluster. Finally, a few tags are suggested to test data using a search-based model and late information fusion strategy.

2.2. Multi-view clustering methods

Thanks to technology development, data can be collected from different sources or described via various feature extraction techniques. According to the complementary information, each view of features might contain helpful knowledge about data that other views do not have. Therefore, multi-view clustering models have attracted a lot of attention over recent decades for exploiting complementary information across different views. Multi-view graph-based clustering methods try to construct a fusion graph across available views and employ different algorithms (e.g., spectral clustering) to cut this fusion graph in order to find the clusters (Cai, Nie, Huang, & Kamangar, 2011; Nie, Cai, Li, & Li, 2018; Nie, Li, & Li, 2016; Nie, Tian, & Li, 2018; Kang et al., 2020; Hu et al., 2020; Cao, Zhang, Fu, Liu, & Zhang, 2015; Zhang, Fu, Liu, Liu, & Cao, 2015; Tang et al., 2019; Wang, Yang, & Liu, 2019). The main point of using graph-based knowledge for dealing with multi-view data is to reasonably fuse the different representations and find the most consistent manifold structure with data distributions. Cai et al. proposed a multi-modal spectral clustering algorithm (MMSC), which calculates different similarity matrices for different modals and then learns a commonly shared Laplacian matrix by integrating different modalities (Cai et al., 2011). Nie et al. provided an efficient algorithm that performs clustering and local structure learning simultaneously (Nie et al., 2018). Their constructed graph can be clustered into specific partitions directly, and the

algorithm allocates ideal weight for each view automatically. Nie et al. also deployed a multi-view model (AMGL), which first constructs the Laplacian matrix for each view (Nie et al., 2016). It then learns an optimal weight for each view automatically and computes the indicator matrix iteratively based on these weights. Moreover, they proposed a multi-view extension of the spectral rotation model in (Nie et al., 2018). Kang et al. provided a new multi-graph fusion method for multi-view spectral clustering (GFSC) to carefully consider the flexible local manifold structure of various views and maintain an explicit cluster structure (Kang et al., 2020). Tang et al. used the low-rank representation (LRR) model to learn a unified similarity graph for multi-view clustering (Tang et al., 2019). They designed a diversity regularization term to learn optimal weights for various views. This term is useful to exploit diversity and reduce the redundancy among different views. Wang et al. proposed a novel graph-based learning method (GMC), in which the similarity graph of each view and the fusion graph are jointly constructed to help each other in a mutual reinforcement manner (Wang et al., 2019).

Although prior multi-view graph-based clustering models have attained significant progress, there are still some limitations. First, most of the existing methods construct a subspace representation for each view, and then the average of these representations is used to find the final clustering result (Cao et al., 2015; Zhang et al., 2015). This approach cannot capture the different contributions of views since it ignores the complementary information across different views. Second, most previous models learn the similarity graph of each view in isolation and keep the learned graph fixed during the fusion step (Nie et al., 2016; Hu et al., 2020). Third, many current methods carry out graph learning and clustering steps separately (Hu et al., 2020; Nie et al., 2016; Cai et al., 2011; Zamiri & Yazdi, 2020). These models do not consider the graphs' quality since they ignore that the fusion graph should have exact k connected components. Furthermore, some methods deal with multiple views indiscriminately (Cao et al., 2015; Zhang et al., 2015). They do not consider the redundancy and diversity among various views. However, the redundancy between similar views should be reduced efficiently while the diversity among views should be appropriately enhanced. Moreover, random noises and outliers frequently exist in real-world data, which degenerate the similarity graph's quality and clustering performance of many current models.

To overcome the aforementioned limitations, we propose a novel multi-view data fusion model using robust spectral clustering, denoted by MVDF-RSC. Our model handles the contributions of different views by automatically generated weights. Also, we learn a fusion similarity graph of various views rather than averaging individual graphs of views constructed in isolation. The cluster number known in advance is utilized for regularizing the fusion similarity graph learning. Therefore, our method does not require an additional clustering step for finding the final clusters. Moreover, we use a diversity regularization term to learn the weights of different views adaptively, which is useful for improving the diversity and suppressing the redundancy among different views. Furthermore, the proposed method is robust against large outliers and noises by providing a MCC-based framework to employ the information in the data efficiently. To the best of our knowledge, no existing method addresses all these five limitations simultaneously. In this paper, we address all these limitations and provide our problem using a robust fusion framework. The proposed method's superiority over state-of-the-art models is validated comprehensively by performing experiments on two multi-view datasets.

2.3. Maximum correntropy criterion

The mean square error (MSE) is the popular methodology for measuring how similar two arbitrary random variables are (AAltan & Karasultan & Karasu, 2019). However, its good performance is reliant on the Gaussianity assumption. The concept of correntropy is proposed (Liu, Pokharel, & Principe, 2007) in ITL to deal with non-Gaussian noises and large outliers. It has been widely applied to many areas, like signal processing (Chen, Xing, Zhao, Zheng, & Principe, 2016; Chen, Xing, Liang, Zheng, & Principe, 2014) and computer vision (Zhou, Xu, Cheng, Yuan, & Chen, 2019; He, Zheng, & Hu, 2011). Correntropy of two random variables X and Y measures how they are similar to each other and is defined as follows:

$$V_\sigma(X,Y) = E[k(X-Y)], \tag{1}$$

where $E[.]$ denotes the mathematical expectation, and $k(.)$ is the kernel function. In practice, since the joint probability density function is usually unknown, for a finite number of data $\{x_i, y_i\}_{i=1}^n$, the correntropy in Eq. (1) can be approximated as follows:

$$\widehat{V_\sigma}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x_i - y_i), \tag{2}$$

where $k_\sigma = g(X,\sigma) = \exp(-X^2/2\sigma^2)$ is the Gaussian kernel with the scaling factor σ. The maximum of Eq. (2) is called the maximum correntropy criterion (MCC) and has been successfully used in different adaptive algorithms robust to large outliers.

3. The proposed MVDF-RSC method

In this section, we propose the MVDF-RSC model in more detail. Generally, there are two types of fusion techniques: early fusion and late fusion.

On the one hand, early fusion methods fuse the features obtained from different views, such as textual features, visual features, etc., as the main feature vector and then analyze these fused feature vectors. Early fusion models' merit is that the correlation between different views at an early step can provide better task accomplishment.

On the other hand, late fusion models examine and classify each view's features independently, and then the results are fused as a decision vector for obtaining the final decision. The fusion of various decisions obtained from different views is easier than the early fusion approach because the output decisions resulting from various views have the same data form.

The proposed method is a hybrid fusion model that uses both early and late fusion strategies to exploit the merits of both of them. As shown in Fig. (2), the proposed method consists of two phases: (1) training phase and (2) prediction phase. In the training phase, features are firstly extracted using different feature extraction techniques. Then, we find the clusters using a new robust multi-view graph-based clustering method and early fusion strategy. Subsequently, landmarks are extracted for each cluster. A few tags are suggested to test data according to the model and late fusion strategy in the prediction phase. Now, let us introduce the proposed model in more detail.

We introduce all notational conventions employed throughout the paper in Table 1. Let assume that there are N_{tr} training images. Input images are illustrated with different feature vectors, i.e., geographical, visual, and textual. Therefore, we have three input matrices for geo-

Table 1
Notation.

Symbol	Description
$\mathbf{x_j}$	The j^{th} column vector of matrix $\mathbf{X} \in R^{d\times n}$
X_{ij}	The $(ij)^{th}$ element of the matrix $\mathbf{X} \in R^{d\times n}$
$Tr(\mathbf{X})$	The trace of the matrix $\mathbf{X}$
$\|\mathbf{x}\|_p$	The l_p-norm of the vector $\mathbf{x}$
$\mathbf{X}^T$	The transpose of the matrix $\mathbf{X} \in R^{d\times n}$
$\mathbf{I}$	The Identity matrix
$\mathbf{1}$	A column vector with all elements one

graphical, visual, and textual views. Let $\mathbf{I_v} \in R^{d_v \times N_{tr}}$ denotes the input feature matrix of v^{th} view, $v = 1, \ldots, V$. d_v denotes the dimension of v^{th} view. For example, in the geographical view, the latitude and longitude of images are employed, so the geographical view dimension is equal to 2. Matrices and vectors are shown in boldface capital letters and boldface lowercase letters respectively. Also, scalars are depicted in lowercase letters.

Given training image data $\{\boldsymbol{I}_{\mathbf{v}}^{\mathbf{i}}\}_{i=1}^{N_{tr}}$ for $v = 1, \ldots, V$, fully-connected graph $\mathbf{W} \in R^{N_{tr} \times N_{tr}}$ shared by different views is constructed by solving the following MCC-based optimization problem:

$$\begin{aligned} &\max_{\mathbf{w_i}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \exp(-\eta_v e_{vi}^2) \quad s.t. \quad \mathbf{1}^T \mathbf{w_i} \\ &= 1, \quad \mathbf{w_i} \geqslant 0, \quad \forall i \\ &= 1, \ldots, N_{tr}, \end{aligned} \tag{3}$$

where $\mathbf{w_i} \in R^{N_{tr}}$ and η_v are the i^{th} column vector of the coefficient matrix $\mathbf{W}$ shared by different views and the scaling factor for v^{th} view respectively. The error for each data sample i and view v, e_{vi}, is defined as follows:

$$\begin{aligned} e_{vi} &= \|\mathbf{G_v^i w_i}\|_2, \quad v \\ &= \{1, 2, \ldots, V\}, \quad i \\ &= \{1, \ldots, N_{tr}\}, \end{aligned} \tag{4}$$

where $\mathbf{G_v^i} \in R^{N_{tr} \times N_{tr}}$ is a diagonal matrix having normalized distances between the i^{th} training data and others based on v^{th} view on the diagonal, and is defined as follows:

$$\mathbf{G_v^i} = Diag(ndist(\mathbf{I_v^1}, \mathbf{I_v^i}), \ldots, ndist(\mathbf{I_v^{N_{tr}}}, \mathbf{I_v^i})), \tag{5}$$

where $\mathbf{I_v^i}$ denotes the i^{th} column of matrix $\mathbf{I_v}$. Since different views may have different scales, we normalize each row of input matrices $\{\mathbf{I_v}\}_{v=1}^{V}$ to have a unit norm of l_2-norm. For example, $ndist(\mathbf{a}, \mathbf{b})$ shows the distance between two arbitrary normalized vectors $\mathbf{a}$ and $\mathbf{b}$. To be specific, the column vector $\mathbf{w_i}$ of the similarity matrix $\mathbf{W}$ in Eq. (3) is constructed by the representation learning model. The representation learning model assumes that each data sample can be reconstructed via a linear combination of other instances, and the reconstruction coefficients, $\{\mathbf{w_i}\}_{i=1}^{N_{tr}}$, show the similarity between samples.

Maximum correntropy is used in Eq. (3) for robustness to large outliers in input data. However, Eq. (3) does not consider the redundancy and diversity among various views. The diversity needs to be efficiently exploited, while the redundancy should be reduced to enhance the clustering results. In this regard, we employ a diversity regularization term like the method proposed by (Tang et al., 2019). Let $\mathbf{T_v}(v = 1, \ldots, V)$ denotes the probability transition matrix for the v^{th} view corresponding to a random walk defined on it. We have $\mathbf{T_v} = \mathbf{D_v^{-1}} \mathbf{S_v}$ where $\mathbf{S_v}$ is the similarity matrix of data for v^{th} view and $\mathbf{D_v}$ is a diagonal degree matrix -$\{D_v\}_{ii} = \sum_{j=1}^{N_{tr}} \{S_v\}_{ij}$. $\mathbf{T_v}(v = 1, \ldots, V)$ has the information of views, and the similarity between two probability transition matrices $\mathbf{T_l}$ and $\mathbf{T_m}$ shows the similarity between the l^{th} and m^{th} views. If the corresponding column/rows of two matrices $\mathbf{T_l}$ and $\mathbf{T_m}$ are highly related, they will be similar to each other, and the inner product of the column/row vectors of them will be large. Thus, the sum of all inner product values, $Tr(\mathbf{T_l^T T_m})$, should be assigned a larger value. As a result, the symmetry matrix $\mathbf{H} \in R^{V \times V}$ measures the similarity between various views, in which the similarity between l^{th} and m^{th} views is obtained as $H_{lm} = H_{ml} = Tr(\mathbf{T_l^T T_m})$. If two views are similar, their corresponding element of the matrix $\mathbf{H}$ will be a large value. Since the scaling factor $\boldsymbol{\eta} \in R_+^V$ indicates the significance of different views for the correntropy function in Eq. (3), we define a diversity regularization term as follows:

$$\min_{\boldsymbol{\eta} \in R_+^V} \sum_{i,j=1}^{V} \eta_i \eta_j H_{ij} = \boldsymbol{\eta}^T \mathbf{H} \boldsymbol{\eta} \quad s.t. \quad \boldsymbol{\eta}^T \mathbf{1_V} = 1, \tag{6}$$

where $\mathbf{1}_V \in R^V$ is a vector in which all elements are 1. In Eq. (6), when i^{th} and j^{th} views are similar, H_{ij} is large, then their corresponding η-s (i.e., η_i and η_j) would not be large simultaneously. Therefore, the scaling factors of the correntropy function control diversity as well as redundancy among different views. Now, we extend Eq. (3) using Eq. (6) to formulate the proposed method as:

$$\begin{aligned} &\max_{\mathbf{w_i}, \boldsymbol{\eta}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \exp\left(-\eta_v \|\mathbf{G_v^i w_i}\|_2^2\right) - \lambda \boldsymbol{\eta}^T \mathbf{H} \boldsymbol{\eta} \quad s.t. \\ &\boldsymbol{\eta}^T \mathbf{1_V} = 1, \mathbf{1}^T \mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \ldots, N_{tr}. \end{aligned} \tag{7}$$

Remark 1 According to Eq. (7), the first term considers the relevance criterion of views to find the optimal $\{\mathbf{w_i}\}_{i=1}^{N_{tr}}$. The optimal scaling factors $\{\boldsymbol{\eta_v}\}_{v=1}^{V}$ are determined using the second term. The parameters $\{\eta_v\}_{v=1}^{V}$ control the error distribution of data. In other words, a larger η_v leads to a smaller $\{e_{vi}\}_{i=1}^{N_{tr}}$. As a result, $\boldsymbol{\eta}$ acts as an error distribution controller, reducing redundancy and exploiting diversity among various views.

The ideal solution $\mathbf{W}$ of the problem (7) is that the data should have exact k connected components, aiming to cluster the data into k clusters. However, the current solution cannot usually reach the ideal condition. This goal can be fulfilled by introducing a rank constraint inspired by the following important property of the Laplacian matrix (Mohar, Alavi, Chartrand, Oellermann, & Schwenk, 1991):

Theorem 1 *The multiplicity k of the zero eigenvalues of the Laplacian matrix* $\mathbf{L_W}$ *is equal to the connected components in the graph with the similarity matrix* $\mathbf{W}$.

According to Theorem 1, if $rank(\mathbf{L_W}) = N_{tr} - k$, then the corresponding similarity matrix $\mathbf{W}$ is our ideal matrix based on which the data are clustered into k partitions directly. Hence, there is no need to apply an additional clustering algorithm on the fusion similarity matrix $\mathbf{W}$ to find the final clusters. By adding a rank constraint into Eq. (7), we formulate our multi-view clustering model as:

$$\begin{aligned} &\max_{\mathbf{w_i}, \boldsymbol{\eta}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \exp\left(-\eta_v \|\mathbf{G_v^i w_i}\|_2^2\right) - \lambda \boldsymbol{\eta}^T \mathbf{H} \boldsymbol{\eta} \\ &s.t. \quad \boldsymbol{\eta}^T \mathbf{1_V} = 1, rank(\mathbf{L_W}) = N_{tr} - k, \mathbf{1}^T \mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \ldots, N_{tr}. \end{aligned} \tag{8}$$

It is hard to solve Eq. (8) since $rank(\mathbf{L_W}) = N_{tr} - k$ is nonlinear and the Laplacian matrix $\mathbf{L_W}$ depends on the variable $\mathbf{W}$.

Let $\sigma_i(\mathbf{L_W})$ denotes the i^{th} eigenvalue of the Laplacian matrix $\mathbf{L_W}$. It is noted that $\sigma_i(\mathbf{L_W}) \geqslant 0, (i = 1, \ldots, N_{tr})$ because $\mathbf{L_W}$ is a positive semi-definite matrix. Then, the constraint $rank(\mathbf{L_W}) = N_{tr} - k$ can be approximately reached if $\sum_{i=1}^{k} \sigma_i(\mathbf{L_W}) = 0$. So, $\sum_{i=1}^{k} \sigma_i(\mathbf{L_W})$ can be minimized instead to achieve our goal. We can write a following objective function based on Ky Fan's theorem (Fan, 1949):

$$\sum_{i=1}^{k} \sigma_i(\mathbf{L_W}) = \min_{\mathbf{F}, \mathbf{F^T F = I}} Tr(\mathbf{F}^T \mathbf{L_W} \mathbf{F}), \tag{9}$$

where $\mathbf{F} \in R^{N_{tr} \times k}$ is the cluster indicator matrix. Now, by plugging Eq. (9) into Eq. (8), our final objective function can be formulated as:

$$\begin{aligned} &\max_{\mathbf{w_i}, \boldsymbol{\eta}, \mathbf{F}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \exp\left(-\eta_v \|\mathbf{G_v^i w_i}\|_2^2\right) - \lambda \boldsymbol{\eta}^T \mathbf{H} \boldsymbol{\eta} - \gamma Tr\left(\mathbf{F}^T \mathbf{L_W} \mathbf{F}\right) \\ &s.t. \quad \boldsymbol{\eta}^T \mathbf{1_V} = 1, \mathbf{F}^T \mathbf{F} = \mathbf{I}, \mathbf{1}^T \mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \ldots, N_{tr}. \end{aligned} \tag{10}$$

We write below equation based on the conjugate function theory (Boyd & Vandenberghe, 2004):

$$\exp(-\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2) = \sup_{p_{vi}<0}(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi})), \tag{11}$$

where $\phi(p_{vi})$ is a convex function $\phi(p_{vi}) = -p_{vi}\log(-p_{vi}) + p_{vi}, p_{vi}<0$ (see A). Then, the Eq. (10) can be rewritten as:

$$\begin{aligned} J(\mathbf{W},\boldsymbol{\eta},\mathbf{F}) &= \textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\sup(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi})) - \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta} - \gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F}),\\ &= \sup_{p_{vi}}\left\{\textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi}))\right\} - \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta} - \gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F}),\\ &= \sup_{p_{vi}}\left\{\textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi})) - \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta} - \gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F})\right\}. \end{aligned} \tag{12}$$

The second equation in the Eq. (12) establishes because $\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi})$, $(i = 1,\dots,N_{tr})$ are independent functions in terms of p_{vi}. Also, the third equation in the Eq. (12) establishes since $-\lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}$ and $-\gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F})$ are constant with respect to p_{vi}. Using Eq. (12), Eq. (10) is equivalent to:

$$\begin{aligned} &\max_{\mathbf{w_i},\boldsymbol{\eta},\mathbf{F},\mathbf{p_v}} J(\mathbf{W},\boldsymbol{\eta},\mathbf{F},\mathbf{p_v}) = \max_{\mathbf{w_i},\boldsymbol{\eta},\mathbf{F},\mathbf{p_v}} \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\left(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \phi(p_{vi})\right) - \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}\\ &s.t. \quad \boldsymbol{\eta}^T\mathbf{1_V} = 1, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{1}^T\mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1,\dots,N_{tr}. \end{aligned}$$

The variables in Eq. (13) are coupled to each other; therefore, solving it to find an optimal solution for every variable at once is difficult. So, we should solve it iteratively, in which each of the parameters is alternately updated while keeping others as constant values.

W-subproblem: By omitting irrelevant variables, the variable $\mathbf{w_i}$ can be obtained by solving the following problem:

$$\begin{aligned} &\max_{\mathbf{w_i}} \textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F})\\ &s.t. \quad \mathbf{1}^T\mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1,\dots,N_{tr}. \end{aligned} \tag{14}$$

By defining $q_{ji} = -p_{ji}$, $j = 1,\dots,V$, $i = 1,\dots,N_{tr}$, Eq. (14) becomes

$$\begin{aligned} &\min_{\mathbf{w_i}} \textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 q_{vi} + \gamma Tr(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F})\\ &s.t. \quad \mathbf{1}^T\mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1,\dots,N_{tr}. \end{aligned} \tag{15}$$

For solving subproblem (15), we use below equality:

$$\sum_{i,j=1}^{N_{tr}}\frac{1}{2}\|\mathbf{F}_{\mathbf{i},:} - \mathbf{F}_{\mathbf{j},:}\|^2 W_{ij} = Tr\left(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F}\right), \tag{16}$$

and we define $\mathbf{h_i} \in R^{N_{tr}}$, which its j^{th} entry is $h_{ij} = \|\mathbf{F}_{\mathbf{i},:} - \mathbf{F}_{\mathbf{j},:}\|^2$. Then, Eq. (15) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{w_i}} \textstyle\sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\eta_v\mathbf{w}_\mathbf{i}^T(\mathbf{G}_\mathbf{v}^{\mathbf{i}^T}\mathbf{G}_\mathbf{v}^\mathbf{i}q_{vi})\mathbf{w_i} + \frac{\gamma}{2}\mathbf{h}_\mathbf{i}^T\mathbf{w_i}\\ &s.t. \quad \mathbf{1}^T\mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1,\dots,N_{tr}. \end{aligned} \tag{17}$$

Eq. (17) is a standard quadratic programming problem with both equality and inequality constraints. We can solve it using the MATLAB toolbox easily.

p-subproblem: We update p_{vi} and fix the other variables, and our optimization problem (13) becomes

$$\max_{p_{vi}} \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} + p_{vi}\log(-p_{vi}) - p_{vi}). \tag{18}$$

By taking the derivative of Eq. (18) with respect to p_{vi} and setting it to zero

$$p_{vi} = -\exp(-\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2). \tag{19}$$

η-subproblem: To update variable $\boldsymbol{\eta}$, we should solve the following problem:

$$\max_{\boldsymbol{\eta_v}} \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}(\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 p_{vi} - \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}) \quad s.t. \quad \boldsymbol{\eta}^T\mathbf{1_V} = 1. \tag{20}$$

By replacing q_{vi} instead of p_{vi}:

$$\begin{aligned} &\min_{\boldsymbol{\eta_v}} \quad \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}\\ &+ \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}}\eta_v\|\mathbf{G}_\mathbf{v}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 q_{vi} \quad s.t. \quad \boldsymbol{\eta}^T\mathbf{1_V} = 1. \end{aligned} \tag{21}$$

We can rewrite Eq. (21) as the following form:

$$\begin{aligned} &\min_{\boldsymbol{\eta_v}} \quad \lambda\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta} + \boldsymbol{\eta}^T\begin{pmatrix}\sum_{i=1}^{N_{tr}}\|\mathbf{G}_\mathbf{1}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 q_{1i}\\ \vdots\\ \sum_{i=1}^{N_{tr}}\|\mathbf{G}_\mathbf{V}^\mathbf{i}\mathbf{w}_\mathbf{i}\|_2^2 q_{Vi}\end{pmatrix} \quad s.t. \quad \boldsymbol{\eta}^T\mathbf{1_V}\\ &= 1. \end{aligned} \tag{22}$$

Eq. (22) is also a standard quadratic programming problem with equality constraint and can be solved using the MATLAB toolbox.

***F*-subproblem**: We update $\mathbf{F}$ by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{F}} \quad \gamma Tr\left(\mathbf{F}^T\mathbf{L}_\mathbf{W}\mathbf{F}\right)\\ &s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}. \end{aligned} \tag{23}$$

The optimal $\mathbf{F}$ can be obtained by k eigenvectors of the Laplacian matrix $\mathbf{L}_\mathbf{W}$ corresponding to the k smallest eigenvalues. The similarity graph learning steps are summarized in Algorithm 1.

Algorithm 1. Similarity Graph Learning

Require: Data for V views $\mathbf{I_1},\dots,\mathbf{I_V}$ with $\mathbf{I_v} \in R^{d_v\times N_{tr}}$, parameters λ and γ
Ensure: Similarity matrix $\mathbf{W}$, $\mathbf{F} \in R^{N_{tr}\times k}$
Initialize: $q_{ji} = 1$, $j = 1,\dots,V$, $i = 1,\dots,N_{tr}$, random matrix $\mathbf{F}$, random vector $\boldsymbol{\eta}$;
for *iter* = 1 to *maxIter* **do**
Update $\mathbf{w_i}$ by solving Eq. (17) $i = 1,\dots,N_{tr}$;
Symmetrize similarity matrix $\mathbf{W_{sym}} = \max(\mathbf{W},\mathbf{W}^T)$;
Update p_{ji} according to Eq. (19), $i = 1,\dots,N_{tr}$, $j = 1,\dots,V$;
Update $\boldsymbol{\eta_v}$ by solving Eq. (22), $v = 1,\dots,V$;
Update $\mathbf{F}$ by solving Eq. (23).
end for

3.1. Landmark extraction via MVDF-RSC Algorithm

Spectral clustering is a well-known clustering method, which determines the clusters based on graph partitioning (Luxburg, 2007; Bao, Guo, & Chai, 2009; Shi & Malik, 2000). According to Eq. (17), the constructed similarity matrix is an asymmetric graph. The below equation is exploited to symmetrize it (Elhamifar & Vidal, 2011):

$$\mathbf{W}_{sym} = \max(\mathbf{W},\mathbf{W}^T). \tag{24}$$

Fig. (3) depicts an example of a similarity matrix constructed through an optimization problem step with fusing geo-tags, visual and textual features and representing the edges' weights. For instance, W_{15} will be greater than W_{16} because of the similarity of visual and textual features between images 1 and 5.

Next, according to the above-constructed similarity matrix ($\mathbf{W}_{sym}$) obtained by using the fusion of different views, the landmarks will be extracted for each cluster. Landmarks are the most repetitive tags in

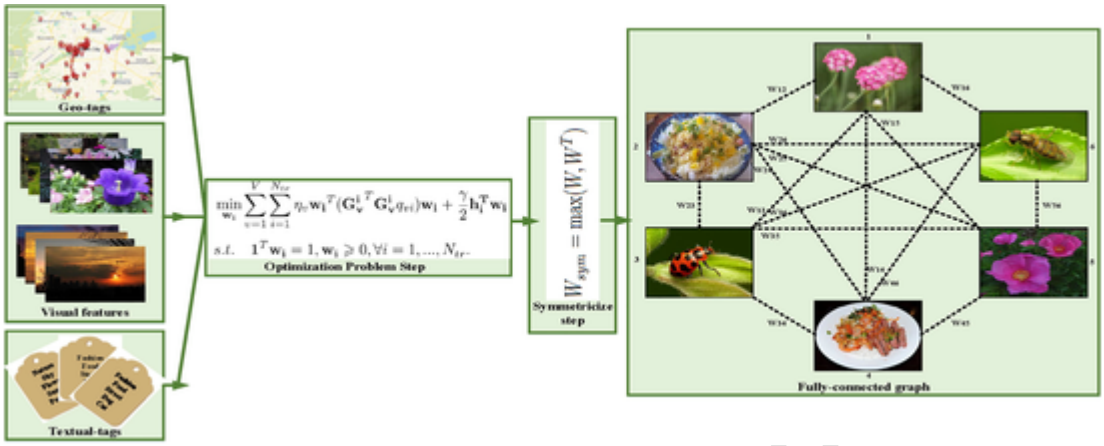

Fig. 3. An overview of the similarity matrix W_{sym} construction.

each cluster. For each cluster, we rank and obtain the landmarks which have the most repetition times.

3.2. Prediction phase: assigning relevant tags to test data

In this phase, the proposed method assigns suitable tags based on a late fusion approach for each unknown test sample. According to Fig. (2), different strategies are implemented to find suitable tags. Euclidean distance is applied to measure the distance between clusters' centers and test data.

- **Geographical strategy:** In this way, the geographical features of test data are compared with the clusters' centers' geographical features. Then, the selected landmarks of the nearest cluster are assigned to it.
- **Visual strategy:** In this approach, the visual distance between test data and clusters' centers is calculated, and the selected landmarks of the nearest cluster are assigned to test data.
- **Geo-Visual strategy:** With concatenating the geographical and visual features, a new vector for each test data and clusters' centers is constructed. Then, we compute the distance of two constructed vectors and assign the selected landmarks of the most similar cluster.
- **Late fusion strategy:** According to the above strategies and their recommended tags, a popular fusion method is used.
  Majority vote (Lam & Suen, 1997; Alizadeh et al., 2019): we assume that each above strategy outputs the nearest cluster's tags. In the majority vote, the "t" tags with the highest vote from all strategies are declared the final recommendation.

The proposed tag recommendation method is summarized in Algorithm 2.

Algorithm 2. Image Tagging using MVDF-RSC

Require: Training data for V views $\mathbf{I}_1, \ldots, \mathbf{I}_V$ with $\mathbf{I}_v \in R^{d_v \times N_{tr}}$, test data $\mathbf{I}_v^{test} \in R^{d_v \times N_{test}}$, N_{test} and t are the number of test data and suggested tags respectively.
Ensure: "t" candidate tags for each test image.
Step 1: Construct similarity graph and find k clusters using Algorithm 1;
for $c = 1$ to k **do**
Step 2: Extract representative concepts (landmarks) of cluster(c) according to Section (3.1);
end for
for $i = 1$ to N_{test} **do**
Step 3: Compute the distance of $I^{test}(i)$ and clusters' centers according to distances defined in Section (3.2);
Step 4: Assign "t" representative concepts of the closest cluster to $I^{test}(i)$ based on the late fusion strategy in the Section (3.2);
end for

3.3. Computational complexity analysis

In Algorithm 2, there are four steps, which its main computational complexity is in step 1 related to the similarity graph construction using Algorithm 1. In Algorithm 1, the computational complexity in solving Eq. (13) consists of four subproblems, i.e., solving $\mathbf{W}, \mathbf{p_v}, \boldsymbol{\eta}$, and $\mathbf{F}$. The main complexity for updating $\mathbf{W}$ and $\mathbf{p_v}$ are $O(N_{tr}^2 V)$ and $O(N_{tr} V)$ respectively. The complexity of solving $\boldsymbol{\eta}$-subproblem is $O(N_{tr} V)$. Updating $\mathbf{F}$ requires calculating the Eigen-decomposition of the Laplacian matrix $\mathbf{L_W}$, which costs $O(N^3)$. Totally, the computational complexity of Algorithm (1) is $O(\text{maxIter}(N_{tr}^2 V + 2N_{tr} V + N^3))$.

3.4. Convergence analysis

Finding a globally optimal solution for our optimization problem in Eq. (13) is still an open problem because Eq. (13) is not a joint convex problem of all variables. We solve this equation using an alternating algorithm, in which each subproblem is convex, and the convergence of each subproblem can be theoretically guaranteed. In the following, we show the convergence of each subproblem.

W-subproblem: Solving $\mathbf{W}$ is a standard quadratic programming problem. Therefore, it has a closed-form solution.

p-subproblem: The maximization objective function in Eq. (18) is a concave function since the second-order derivative of it with respect to p_{vi} is $\frac{1}{p_{vi}}, p_{vi}<0$ (negative value). Hence, the minimization of this objective function is a convex function and decreases monotonically.

η-subproblem: The optimization problem for updating η is also a quadratic programming problem, and it has a closed-form solution.

F-subproblem: The Hessian matrix of Eq. (23) is:

$$\frac{\partial^2 Tr(\mathbf{F}^T \mathbf{L_W} \mathbf{F})}{\partial \mathbf{F} \partial \mathbf{F}^T} = \mathbf{L_W} + \mathbf{L_W}^T, \tag{25}$$

since the Laplacian matrix $\mathbf{L_W}$ is positive semi-definite, the Hessian matrix is also positive semi-definite. Therefore, Eq. (23) is a convex function with respect to $\mathbf{F}$.

4. Experimental result

Extensive experiments are carried out on two geo-tagged image datasets to evaluate the proposed method's effectiveness.

4.1. Datasets

In this paper, experiments are carried out over two geo-tagged image datasets, i.e., Flickr[4] and 500PX.[5] They contain 7,000 geo-tagged images collected by the students of the Pattern Recognition Laboratory of the Ferdowsi University of Mashhad. Flickr and 500PX images have six and eight tags on average respectively. Images were uploaded by users publicly. Example images of these datasets and their tags are shown in Fig. 4. Table 2 provides useful information about them.

4.2. Feature extraction phase

In this section, we elaborate on the feature extraction of datasets in more detail.

4.2.1. Visual feature extraction using convolutional neural network (CNN)

We employ the AlexNet model (Krizhevsky, Sutskever, & Hinton, 2012) trained on the ImageNet dataset (Deng et al., 2009) to extract visual features of images. In this way, for each image, a front-end computation is implemented, and the output of fully connected 7 (fc7) in layer 18 is extracted as visual features.

4.2.2. Textual feature extraction using term frequency-inverse document frequency (TF-IDF)

To extract textual features, we use the TF-IDF method (Salton & Buckley, 1988). TF-IDF is a popular numerical statistic method in Information Retrieval (IR) and Natural Language Processing (NLP), which indicates the importance of a word in a document and is calculated as follows:

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \tag{26}$$

where $tf_{i,j}$, df_i and N represent the number of occurrences of i^{th} tag in tags of j^{th} image, the number of images that contain i^{th} tag and total number of images respectively.

4.2.3. Geographical feature extraction

From photo-sharing websites (i.e., http://www.Flickr.com and http://www.500px.com), we gathered geo-tagged images with latitude and longitude information, and we use them as useful information about the content of images.

4.3. Evaluation criteria

Standard metrics in information retrieval (i.e., Average-Precision (APR), Average-Recall (ARC) and F-measure (F1)) are utilized to evaluate the performance of the proposed method and compare it with different baseline algorithms:

$$\begin{aligned} APR &= \frac{1}{N_{test}}\sum_{i=1}^{N_{test}}\frac{T_k \cap GT}{T_k}, \\ ARC &= \frac{1}{N_{test}}\sum_{i=1}^{N_{test}}\frac{T_k \cap GT}{GT}, \\ F1 &= \frac{2\times APR\times ARC}{APR+ARC}, \end{aligned} \tag{27}$$

where N_{test} denotes the number of test images, T_K is the set of top-K tags returned by the method, and GT denotes the ground-truth tags.

[4] http://www.Flickr.com.
[5] http://www.500px.com.

4.4. Comparison algorithms

We evaluate the effectiveness of the proposed MVDF-RSC with both single-view and multi-view clustering methods.

- Spectral Clustering (SC) (Shi & Malik, 2000): We use conventional spectral clustering as a baseline model. SC(geo), SC(visual), and SC(CF) represent the implementation of SC on geo feature matrix, visual feature matrix, and concatenated features (CF) of all views respectively.
- K-means clustering (KM): We carry out k-means on the concatenated features. It is assumed that visual, textual, and geographical views have the same contributions to the clustering task.
- Auto-weighted multiple graph learning (AMGL) (Nie et al., 2016): This method learns a similarity graph via an adaptive neighbors strategy.
- Multi-view learning with adaptive neighbors (MLAN) (Nie et al., 2018): This SC-based method conducts clustering and local structure learning simultaneously.
- Multi-modal spectral clustering (MMSC) (Cai et al., 2011): It proposes a novel multi-modal spectral clustering to construct a commonly shared similarity matrix by integrating different image modals.
- Multi-view clustering via adaptively weighted procrustes (AWP) (Nie et al., 2018): It provides a new multi-view extension of the spectral rotation model that tries to find an indicator matrix from k spectral embeddings.
- Multi-graph fusion for multi-view spectral clustering (GFSC) (Kang et al., 2020): This method simultaneously conducts graph fusion and spectral clustering.
- The method in (Tang et al., 2019): It employed a low-rank representation model and a diversity regularization term to learn a fusion similarity graph for multi-view subspace clustering.
- GMC (Wang et al., 2019): This method proposed a new learning model, in which graph matrix construction of each view and fusion graph learning help each other in a mutual reinforcement manner.
- MVRSC (Zamiri & Yazdi, 2020): This method provided a robust multi-view graph-based clustering model via MCC-framework. Unlike our method, there is no diversity regularization term and a rank constraint to directly cluster data into k partitions.

Moreover, to evaluate the three views' influences on image tagging, we implement six scenarios and compare their performances. These scenarios are:

- OG: In this scenario, we assume that only the images' geographical view is available. So, we run the MVDF-RSC algorithm using geo-tags and recommend tags to test data based on their locations ($v = \{Geographical\}$).
- OV: In this case, we assume that only visual features are available and implement the proposed model using visual view ($v = \{Visual\}$).
- OGV: In this scenario, we employ both geographical and visual views to find the training phase's fusion similarity matrix. In the prediction phase, it employs the late fusion strategy to suggest relevant tags to test data ($v = \{Visual, Geographical\}$).
- OGT: In this case, we construct the fusion similarity matrix based on the geographical and textual view in the training phase ($v = Geographical, Textual$). Then, relevant tags are recommended based on the geographical strategy in Section (3.2).
- OVT: Visual and textual views are employed to construct the fusion graph ($v = \{Visual, Textual\}$). In the prediction phase, suitable tags are recommended based on the visual strategy in Section (3.2).

Fig. 4. Example images of Flickr (first row) and 500PX (second row) with their corresponding tags.

Table 2
Information about Flickr and 500PX datasets (#Feature).

#View	Flickr	500PX
1	Geographical (2)	Geographical (2)
2	Visual (4096)	Visual (4096)
3	Textual (150)	Textual (117)
#Total image	7000	7000
#Training image (75%)	5250	5250
#Test image (25%)	1750	1750
#Tags per image (on average)	6	8
#Tags	150	117

- GTV (our proposed method): In this scenario, we construct the fusion graph by integrating visual, textual, and geographical views in the training phase. Then, tags are suggested based on the late fusion strategy.

4.5. Parameter setting

Note that GFSC, MLAN, MMSC, GMC, the method in (Tang et al., 2019), and MVRSC use the raw data matrices as input, while AMGL and AWP require the similarity matrices to serve as input. The similarity matrices in AMGL and AWP methods are constructed using the same method with fixed parameters adopted by their authors. The implementation of all the comparison algorithms is publicly available. We tune their parameters by following the parameter settings in their papers for a fair comparison and reporting the results. There are two parameters λ and γ and require to be set properly in the proposed method. We set their values by performing a grid search method on a random subset of training data with N_{param} samples for training and evaluation. We find the optimal value of them by minimizing the average error of all views. The Average-Error is defined as follows:

$$Average - Error = \frac{1}{N_{param} * |v|} \sum_{j=1}^{V} \sum_{i=1}^{N_{param}} e_{ji}, \tag{28}$$

where $|v|$ denotes the number of available views. We report the parameter effect on the Flickr and 500PX datasets in Fig. 5. It is clear that the optimal λ and γ are $\{0.001, 0.01\}$ and $\{10, 0.001\}$ for the Flickr and 500PX respectively.

4.6. Comparison results and analysis

In this section, we evaluate the effectiveness of the proposed method by performing extensive experiments on two geo-tagged datasets. First, we run the proposed algorithm with various scenarios explained in Section (4.4). Second, we compare the MVDF-RSC with other state-of-the-art multi-view graph-based clustering algorithms. Third, we examine the performance of different strategies used in the prediction phase. Moreover, we show the superiority of our algorithm by providing various figures.

The performance of the six scenarios is compared in Table 3. The GTV scenario outperforms other modalities, which confirms that fusing geographical, visual, and textual views in both training and prediction phases helps image tagging tasks. Also, comparing multi-view scenarios (i.e., OGT, OGV, OVT, GTV) with single-view scenarios (i.e., OG and OV), it is obvious that multi-view scenarios show better performance than single-view ones. The OGV achieves better performance than OVT and OGT since this model uses both early and late fusion strategies based on visual and geographical views in the training and prediction

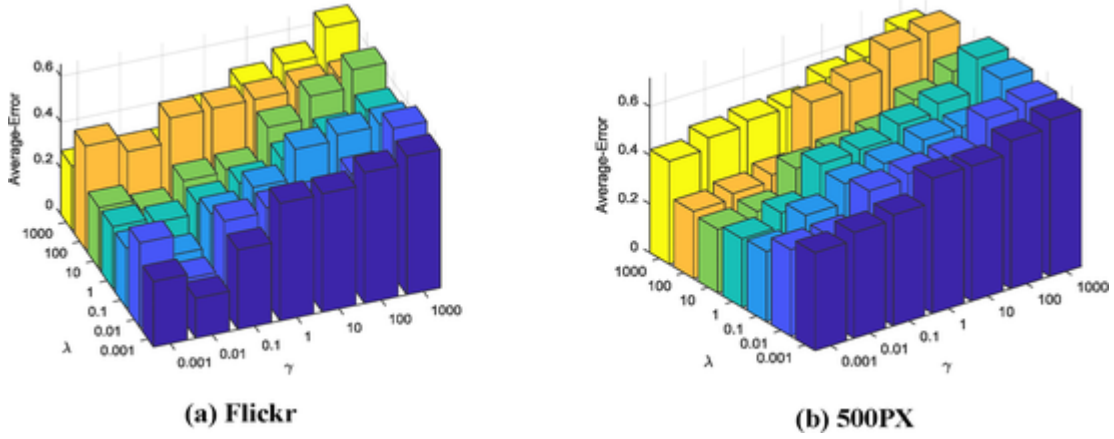

Fig. 5. Average-Error with the variation of λ and γ on the Flickr and 500PX datasets.

Table 3
Recall, Precision, and F1 for different scenarios on the Flickr and 500PX datasets (%).

	Flickr			500PX		
Method	APR	ARC	F1	APR	ARC	F1
OG	17.51	20.35	18.82	13.82	16.22	14.92
OV	19.48	24.51	21.71	16.75	18.11	17.40
OGT	22.61	27.81	24.94	16.73	18.82	17.71
OGV	43.53	43.81	43.67	31.53	35.40	33.35
OVT	23.62	28.21	25.71	19.26	20.91	20.05
GTV	49.69	50.84	50.26	32.07	36.14	33.98

phases. However, OVT and OGT scenarios employ only early fusion techniques in their training phase.

Table 4 compares the performance of image tagging with different multi-view clustering methods and the proposed MVDF-RSC on Flickr and 500PX datasets, in which the best results are marked in boldface. We can draw several conclusions as:

- The performance of conventional SC on different views is compared and each view produces a unique result. This comparison confirms the heterogeneity of visual and geographical views. As a result, it is necessary to distinguish views and consider the redundancy and diversity for building a multi-view clustering algorithm, as we propose in this paper.
- Comparing SCs with the proposed method, it is clear that our proposed MVDF-RSC model achieves better performance since we construct a more accurate and robust similarity graph in our model. Note that we use MCC-based graph learning and a diversity regularization term for weighting different views adaptively. Also, we employ both early and late fusion techniques to employ the advantages of both fusion techniques.
- All multi-view clustering methods achieve better performance than single-view spectral clustering models, confirming that combining various views can enhance the clustering performance.
- Comparing SC(geo) and SC(visual) in Table 4 with OG and OV in the Table 3 shows that OG and OV scenarios outperform SC(geo) and SC(visual), confirming that the proposed graph learning algorithm achieves better results than conventional spectral clustering since the proposed method conducts graph learning and spectral clustering simultaneously.

Table 4
Clustering Methods' Performance on Flickr and 500PX datasets (%).

	Flickr			500PX		
Method	APR	ARC	F1	APR	ARC	F1
SC(geo)	17.61	18.71	18.14	13.33	16.14	14.6
SC(visual)	18.95	22.74	20.67	16.09	17.92	16.96
SC(CF)	20.07	21.35	20.69	18.64	18.01	18.32
KM(CF)	19.72	20.11	19.91	17.06	18.97	17.96
AMGL	26.14	36.92	30.61	17.09	20.74	18.74
MLAN	26.45	45.43	33.43	19.14	23.64	21.15
MMSC	28.73	37.33	32.47	17.93	21.10	19.39
AWP	32.81	40.67	36.32	23.25	27.70	25.28
GFSC	40.21	43.72	41.89	27.93	30.11	28.98
Method in (Tang et al., 2019)	35.90	45.12	39.98	23.82	26.34	25.02
GMC	39.31	43.57	41.33	29.70	31.81	30.72
MVRSC	44.03	45.44	44.7	**32.13**	35.21	33.60
MVDF-RSC	49.69	50.84	50.26	32.07	36.14	33.98

- The proposed multi-view clustering method noticeably outperforms other multi-view clustering models. It shows the superiority of the proposed multi-view graph construction for the spectral clustering algorithm through a robust measure compared with state-of-the-art multi-view spectral clustering methods.
- Compared with MVRSC, the proposed method has achieved significant image tagging progress since it considers diversity and redundancy of views and promotes the similarity graph learning by adding a rank constraint to the objective function.

Table 5 shows the proposed method's results with different strategies introduced in the prediction phase. It is obvious that the fusion of different views in the tagging process achieves a significant performance over other strategies on both Flickr and 500PX datasets. The proposed late fusion technique has achieved the value of $F1 = 50.26\%$ and $F1 = 33.98\%$ over all tags of Flickr and 500PX datasets respectively. Furthermore, geo-visual strategy (i.e., concatenating visual and geographical features) has achieved better performance than using single-view strategies (geographical or visual). Moreover, geographical strategy outperforms visual strategy for the Flickr dataset, proving that geographical view is a good discriminator feature for image tagging.

Figs. 6 and 7 show the impact of tags' popularity on image tagging task for Flickr and 500PX datasets respectively. These tags are sorted in ascending order of $F1$. For example, the proposed method has achieved $F1 = 91\%$ and $F1 = 85\%$ for tag "louisiana" of Flickr dataset and "winter" of 500PX dataset respectively. It is clear that there is a relationship between the popularity of tags and the performance of the proposed method.

The influence of different sizes of training data on the Flickr dataset is considered in Fig. 8. The proposed method has achieved the average value of $F1 = 0.50\%$ over all tags in training data size 5500. This chart illustrates that APR, ARC, and F1 increase when the number of data instances in the training phase increases.

Precision-recall curves for six query tags (i.e., flower, food, ottawa, quebec, wildflower, wildlife) are depicted in Fig. 9. It can be observed that MVDF-RSC yields higher values in precision and recall and a better overall image retrieval performance than MVRSC (Zamiri & Yazdi, 2020) and Geo-kmeans model (Abbasi, Grzegorzek, & Staab, 2009) because the proposed method promotes the graph construction by adding a rank constraint and considers diversity and redundancy criteria. Also, the proposed method can recommend geographical tags (indirectly content-related tags) such as "quebec" and "ottawa" to images as well as directly content-related tags with satisfactory accuracy.

Figs. 10 and 11 show several random-selected test images from the Flickr dataset labeled with "beach" and "garden" tags by the proposed method. Similarly, Figs. 12 and 13 illustrate some randomly selected images from the 500PX dataset and are labeled with "food" and "cold" tags by the proposed method. These figures approve the satisfactory performance of the proposed method to assign relevant tags to test images.

In Fig. 14, we randomly selected some test images with their ground truth tags from Flickr and 500PX datasets. We compare tags

Table 5
Experimental results of the proposed MVDF-RSC on two datasets, Flickr and 500PX, according to strategies in the prediction phase (see Section 3.2).

	Flickr			500PX		
Strategy	APR	ARC	F1	APR	ARC	F1
Geographical	41.70	42.88	42.28	16.53	19.30	17.81
Visual	33.75	40.90	36.98	20.05	22.02	20.99
Geo-Visual	41.52	46.05	43.66	31.59	33.14	32.35
Late Fusion	49.69	50.84	50.26	32.07	36.14	33.98

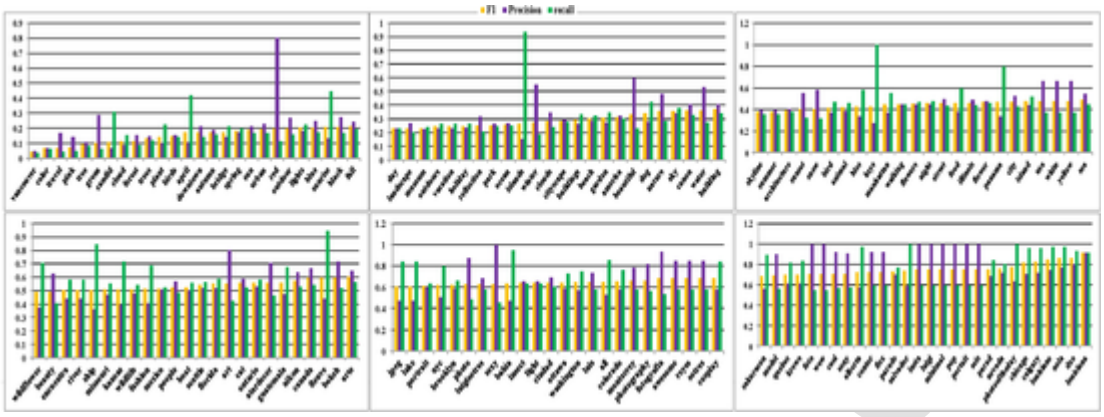

Fig. 6. The performance of the image tagging method on the Flickr dataset for different tags.

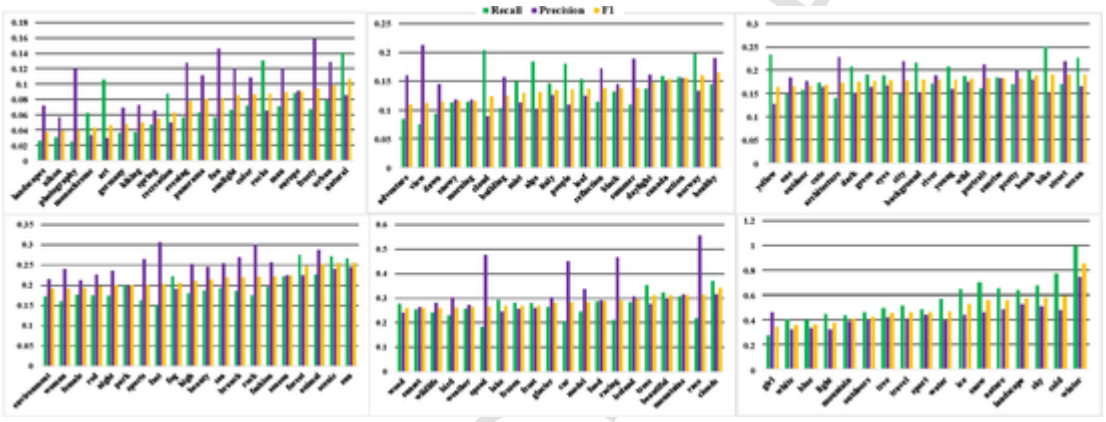

Fig. 7. The performance of image tagging method on 500PX dataset for different tags.

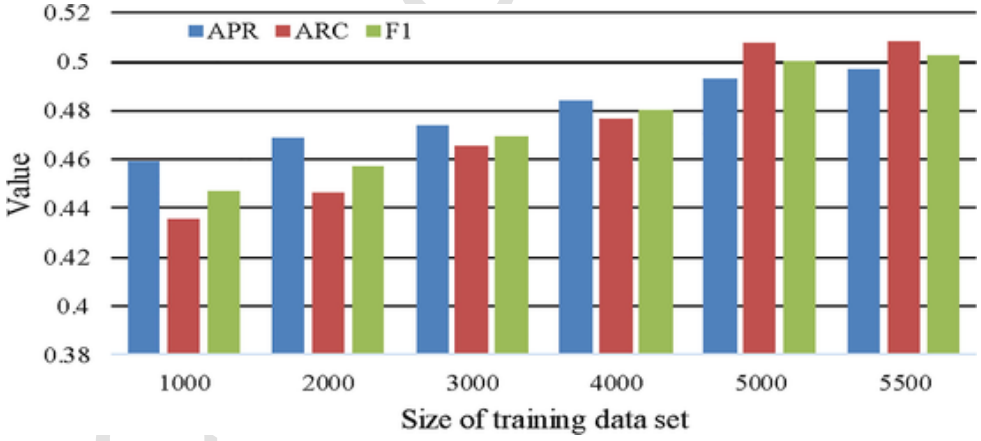

Fig. 8. Influences of different sizes of the training data samples on the proposed method for Flickr dataset.

recommended by the proposed method and MVRSC (Zamiri & Yazdi, 2020). It is obvious that MVDF-RSC suggests more relevant tags in most images than MVRSC. In some examples (i.e., (row 1, col 5), and (row 2 col4)), the proposed method annotated tags which are matched with the ground truth completely. However, in some images, some tags are not matched with the ground truth but describe the image's content. For instance, "sky" and "frozen" are appropriate tags with the content of images (row1, col 4) and (row 2, col 2) respectively. This problem is because of lacking ideal ground truth tags.

Remark 2 In Fig. 14, compared with MVRSC, the proposed method tries to suggest more diverse tags to the images to cover various aspects of target images. The image in (row 2, col 1), for example, there are redundant "sunrise" and "sunlight" among tags recommended by MVRSC while the proposed method retrieved the tag "landscape" instead. This figure shows that the proposed method tries to find more diverse and relevant tags for images than MVRSC.

Figs. 15(a) and 16(a) show the most representative images for two geo-tags ("Washington" and "Bahia") of the Flickr dataset respectively. Moreover, the "Washington" and "Bahia" cluster distribution functions are shown in Figs. 15(b) and 16(b) respectively. The distribution function describes the geographical locations (i.e., latitude and longitude) where photos are taken there. There are peaks in locations

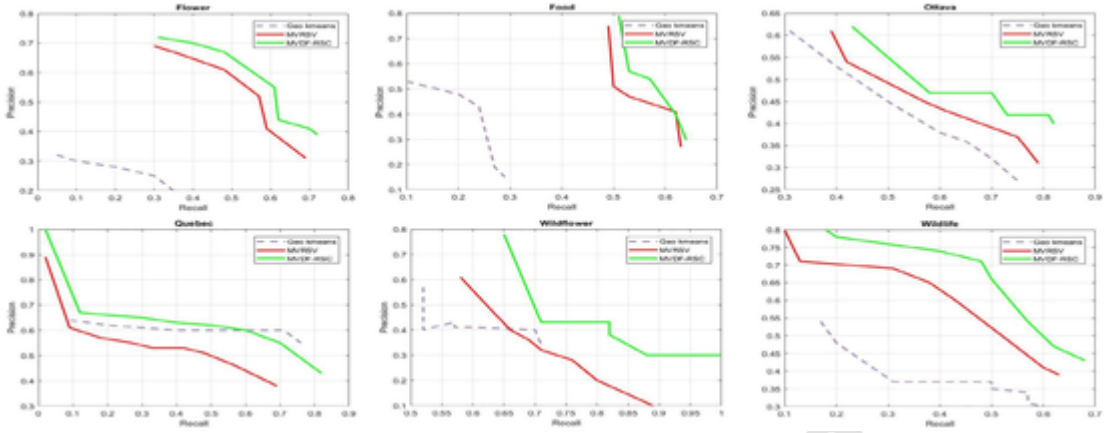

Fig. 9. Precision-recall curves for six query tags on the Flickr dataset for the proposed method, MVRSC (Zamiri & Yazdi, 2020), and Geo-kmeans model (Abbasi et al., 2009).

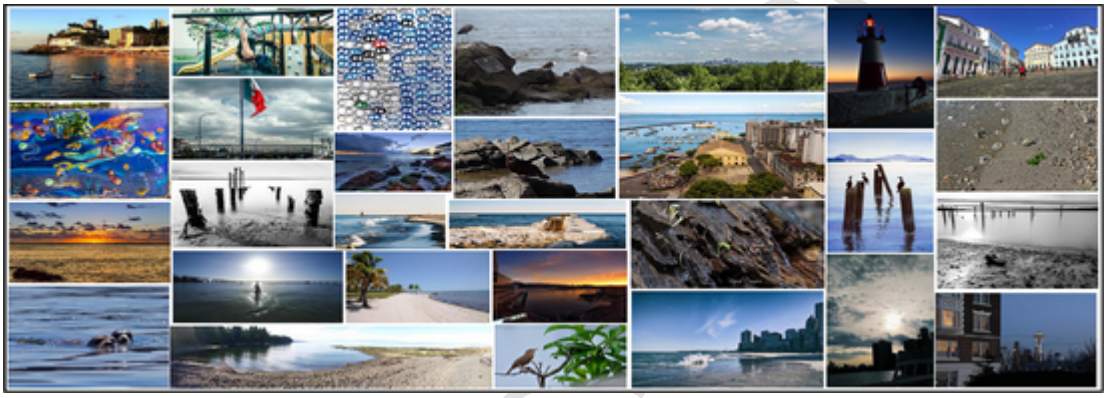

Fig. 10. The proposed method results for several random-selected test images for cluster "beach" of the Flickr dataset.

Fig. 11. The proposed method results for some random-selected test images for cluster "garden" of the Flickr dataset.

where more images are shared. These maps[6] confirm the effectiveness of the proposed method on recommending appropriate geo-tags to images.

5. Conclusion

In this paper, we proposed a new multi-view robust spectral clustering method for image tag recommendation. Since multimedia content is user-generated, it may be corrupted by noise. We provide a robust model against large outliers using the maximum correntropy criterion. Moreover, the correspondence between geospatial information and visual features are used, helping to find indirectly content-related tags. Also, unlike many current image tagging models, which often assign tags to the target images based on the relevance criterion, we propose a way to consider not only the relevance criterion but also diversity and redundancy among different views. In this regard, we use a diversity regularization term to enforce the diversity and suppress the redundancy of various views. Furthermore, the cluster structure of the fusion similarity graph is considered with a rank constraint. Experiments on two geo-tagged datasets verify the effectiveness of the proposed method. Several viable future works can be listed as follows:

- We plan to examine other robust loss functions in the proposed framework.

[6] Maps are visualized using eSpatial website (https://maps.espatial.com).

Fig. 12. The proposed method results for several random-selected test images for cluster "food" of the 500PX dataset.

Fig. 13. The proposed method results for some random-selected test images for cluster "cold" of the 500PX dataset.

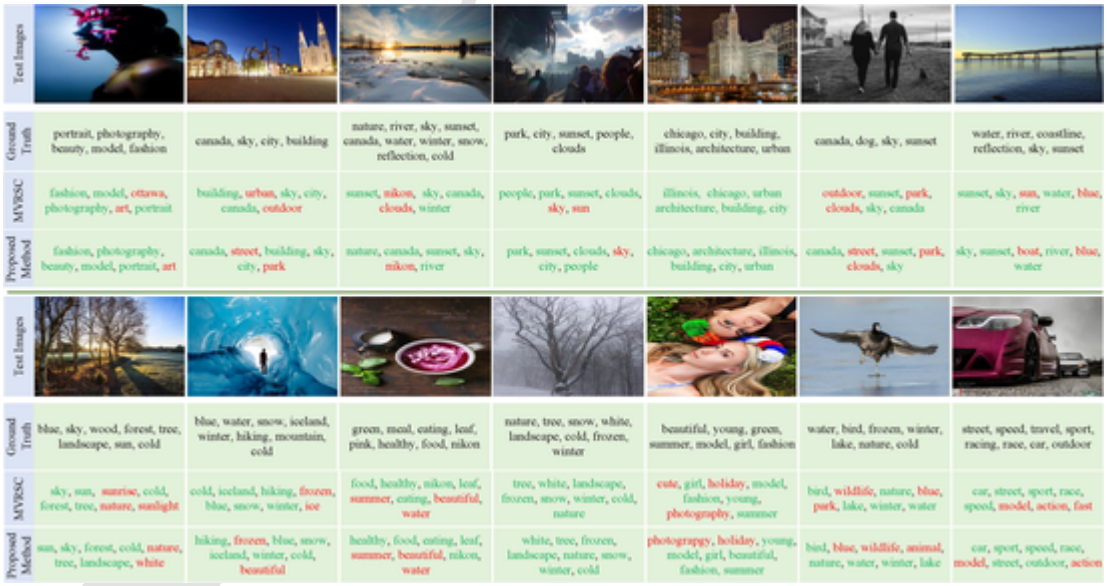

Fig. 14. Tagging examples on the Flickr and 500PX datasets. The first and second rows of images show the examples from Flickr and 500PX datasets respectively. Ground truth tags are those in black, and the tags in color are suggested by the proposed MVDF-RSC method and MVRSC (Zamiri & Yazdi, 2020) in which green words are matched tags, and red ones are mismatching tags in the ground truth.

- We can adjust a weighting scheme to find relevant tags for test data in the prediction phase to distinguish the views' contributions.
- We will try to change the fixed number of tags' strategy for a better prediction.

CRediT authorship contribution statement

Mona Zamiri: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Tahereh Bahraini:** Conceptualization, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Hadi Sadoghi Yazdi:** Conceptualization, Supervision, Writing - review & editing.

Fig. 15. The proposed method results for Flickr dataset (a) several random-selected test images for cluster "washington" and (b) a map view of test image distribution which "washington" is assigned to them. There are peaks in locations where more images are shared.

Fig. 16. The proposed method results for Flickr dataset (a) some random-selected test images for cluster "bahia" and (b) a map view of test image distribution which "bahia" is assigned to them. There are peaks in locations where more images are shared.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Half-quadratic (HQ) optimization

In this section, we prove Eq. (11) in more detail. The description of HQ modeling is stated based on the conjugate function theory (Boyd & Vandenberghe, 2004). The conjugate function $\phi^*(s)$ for a convex function $\phi(p)$ is defined as

$$\phi^*(s) = \sup_{p \in dom\phi} (s^T p - \phi(p)), \tag{29}$$

where $\phi(p) = -p\log(-p) + p, p<0$ can be defined as a convex function. Then, by plugging it to Eq. (29):

$$\phi^*(s) = \sup_{p \in dom\phi} (s^T p + p\log(-p) - p). \tag{30}$$

The above equation can be solved by setting the derivative of $s^T p + p\log(-p) - p$ to 0 since the function $\phi^*(s)$ is concave with respect to p. So, we have:

$$\begin{aligned} &\frac{\partial\{s^T p + p\log(-p) - p\}}{\partial p} = 0 \\ &\rightarrow s + \log(-p) = 0 \\ &\rightarrow p = -\exp(-s) < 0. \end{aligned} \tag{31}$$

As a result, the supremum can be found at $p = -\exp(-s)$. By substituting p in Eq. (30), the conjugate function of $\phi(p)$ is $\phi^*(s) = \exp(-s)$.

References

Abbasi, R., Grzegorzek, M., & Staab, S. (2009). Large scale tag recommendation using different image representations. International Conference on Semantic and Digital Media Technologies, 5887, 65–76.

Alizadeh, R., Jia, L., Nellippallil, A.B., Wang, G., Hao, J., Allen, J.K., & Mistree, F. (2019). Ensemble of surrogates and cross-validation for rapid and accurate predictions using small data sets. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 33, 1–18.

Altan, A., & Karasu, S. (2019). The effect of kernel values in support vector machine to forecasting performance of financial time series and cognitive decision making. Cognitive Systems, 4, 17–21.

Altan, A., & Karasu, S. (2020). Recognition of covid-19 disease from x-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique. Chaos, Solitons & Fractals, 140, 110071.

Amiri, S.H., & Jamzad, M. (2018). Leveraging multi-modal fusion for graph-based image annotation. Visual Communication and Image Representation, 55, 816–828.

Atrey, P.K., Hossain, M.A., Saddik, A.E., & Kankanhalli, M.S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16, 345–379.

Bao, S., Guo, C., & Chai, S. (2009). A note on spectral clustering method based on normalized cut criterion. In *Chinese conference on pattern recognition* (pp. 1–5).

Bar-Hillel, A., Hertz, T., Shental, N., & Weinshall, D. (2005). Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 6, 937–965.

Boulos, M.N.K., Peng, G., & VoPham, T. (2019). An overview of GeoAI applications in health and healthcare. International Journal of Health Geographics, 18, 7.

Boyd, S., & Vandenberghe, L. (2004). Convex optimization. New York: Cambridge University Press.

Cai, G., Lee, K., & Lee, I. (2018). Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos. Expert Systems With Applications, 94, 32–40.

Cai, X., Nie, F., Huang, H., & Kamangar, F. (2011). Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR* (pp. 1977–1984).

Cao, X., Zhang, C., Fu, H., Liu, S., & Zhang, H. (2015). Diversity-induced multi-view subspace clustering. IEEE conference on computer vision and pattern recognition (pp. 586–594).

Chen, B., Xing, L., Liang, J., Zheng, N., & Principe, J.C. (2014). Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. IEEE Signal Processing Letters, 21, 880–884.

Chen, B., Xing, L., Zhao, H., Zheng, N., & Principe, J.C. (2016). Generalized correntropy for robust adaptive filtering. IEEE Transactions on Signal Processing, 64, 3376–3387.

Deldjoo, Y., Schedl, M., Cremonesi, P., & Pasi, G. (2020). Recommender systems leveraging multimedia content. ACM Computing Surveys, 53, 1–38.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Computer Vision and Pattern Recognition, 248–255.

Elhamifar, E., & Vidal, R. (2011). Sparse manifold clustering and embedding. In *Neural information processing systems* (p. 55–63).

Fan, K. (1949). On a theorem of weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of sciences of the United States of America, 35*, 652–655.

Feng, S.L., Manmatha, R., & Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. Computer Vision and Pattern Recognition, 2, 1002–1009.

Ghoshal, A., Ircing, P., & Khudanpur, S.P. (2005). Hidden markov models for automatic annotation and content-based retrieval of images and video. Conference on research and development in information retrieval (pp. 544–551).

He, R., Zheng, W.-S., & Hu, B.-G. (2011). Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33, 1561–1576.
Hu, H., tong Zhou, G., Deng, Z., Liao, Z., & Mori, G. (2016). Learning structured inference neural networks with label relations. IEEE conference on computer vision and pattern recognition (pp. 2960–2968).
Hu, Z., Nie, F., Chang, W., Hao, S., Wang, R., & Li, X. (2020). Multi-view spectral clustering via sparse graph learning. Neurocomputing, 384, 1–10.
Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. Conference on research and development in information retrieval (pp. 119–126).
Jia, R., Khadka, A., & Kimc, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. Accident Analysis and Prevention, 121, 223–230.
Jiang, K., Yin, H., Wang, P., & Yu, N. (2013). Learning from contextual information of geo-tagged web photos to rank personalized tourism attractions. Neurocomputing, 119, 17–25.
Kalayeh, M.M., Idrees, H., & Shah, M. (2014). NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization. IEEE conference on computer vision and pattern recognition (pp. 184–191).
Kang, Z., Shi, G., Huang, S., Chen, W., Pu, X., Zhou, J.T., & Xu, Z. (2020). Multi-graph fusion for multi-view spectral clustering. Knowledge-Based Systems, 189, 102–105.
Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy, 212, 118750.
Kou, N.M., Leong Hou, U., Yang, Y., & Gong, Z. (2015). Travel topic analysis: a mutually reinforcing method for geo-tagged photos. GeoInformatica, 19, 693–721.
Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 1, 1097–1105.
Kuo, C.-L., Chan, T.-C., Fan, I.-C., & Zipf, A. (2018). Efficient method for POI/ROI discovery using flickr geotagged photos. ISPRS International Journal of Geo-Information, 7, 121.
Lam, L., & Suen, C.Y. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics-part A: Systems and Humans, 27, 553–568.
Lee, S. S., Won, D., & McLeod, D. (2008). Tag-geotag correlation in social networks. In *ACM workshop on search in social media* (pp. 59–66).
Lei, C., Liu, D., & Li, W. (2016). Social diffusion analysis with common-interest model for image annotation. IEEE Transactions on Multimedia, 18, 687–701.
Li, Y., Shi, X., Du, C., Liu, Y., & Wen, Y. (2016). Manifold regularized multi view feature selection for social image annotation. Neurocomputing, 204, 135–141.
Li, Z., Shi, Z., Zhao, W., Li, Z., & Tang, Z. (2013). Learning semantic concepts from image database with hybrid generative/discriminative approach. Engineering Applications of Artificial Intelligence, 26, 2143–2152.
Liu, J., Li, Z., Tang, J., Jiang, Y., & Lu, H. (2014). Personalized geo-specific tag recommendation for photos on social websites. IEEE Transactions on Multimedia, 16, 588–600.
Liu, W., Pokharel, P.P., & Principe, J.C. (2007). Correntropy: Properties and applications in non-gaussian signal processing. IEEE Transactions on Signal Processing, 55, 5286–5298.
Logesh, R., Subramaniyaswamy, V., Malathi, D., Sivaramakrishnan, N., & Vijayakumar, V. (2020). Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method. Neural Computing and Applications, 32, 2141–2164.
Luxburg, U.V. (2007). A tutorial on spectral clustering. Statistics and Computing, 17, 395–416.
Makadia, A., Pavlovic, V., & Kumar, S. (2008). A new baseline for image annotation. European Conference on Computer Vision, 5304, 316–329.
Mohar, B., Alavi, E.Y., Chartrand, G., Oellermann, O.R., & Schwenk, A.J. (1991). The laplacian spectrum of graphs. Graph Theory, Combinatorics, 871–898.
Murthy, V.N., Can, E.F., & Manmatha, R. (2014). A hybrid model for automatic image annotation. International conference on multimedia retrieval (pp. 369–376).
Nie, F., Cai, G., Li, J., & Li, X. (2018). Auto-weighted multi-view learning for image clustering and semi-supervised classification. IEEE Transactions on Image Processing, 27, 1501–1511.
Nie, F., Li, J., & Li, X. (2016). Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. International Joint Conference on Artificial Intelligence (pp. 1881–1887).
Nie, F., Tian, L., & Li, X. (2018). Multiview clustering via adaptively weighted procrustes. International conference on knowledge discovery and data mining (pp. 2022–2030).
Nwana, A.O., & Chen, T. (2017). Querying users as oracles in tag engines for personalized image tagging. IEEE MultiMedia, 24, 66–75.
Parapar, J., Bellogin, A., Castells, P., & Barreiro, A. (2013). Relevance-based language modelling for recommender systems. Information Processing and Management, 49, 966–980.
Putthividhy, D., Attias, H. T., & Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. *Computer vision and pattern recognition* (pp. 3408–3415).
Qian, X., Liu, X., Zheng, C., Du, Y., & Hou, X. (2013). Tagging photos using users' vocabularies. Neurocomputing, 111, 144–153.
Qian, X., Wu, Y., Li, M., Ren, Y., Jiang, S., & Li, Z. (2020). LAST: Location-appearance-semantic-temporal clustering based POI summarization. IEEE Transactions on Multimedia, 23, 378–390.
Rad, R., & Jamzad, M. (2015). Automatic image annotation by a loosely joint non-negative matrix factorisation. IET Computer Vision, 9, 806–813.
Rad, R., & Jamzad, M. (2017). Image annotation using multi-view non-negative matrix factorization with different number of basis vectors. Visual Communication and Image Representation, 46, 1–12.
Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 24, 513–523.
Savita, P., Patel, D., & Sinhal, A. (2013). A neural network approach to improve the efficiency of image annotation. International Journal of Engineering Research and Technology, 2, 35–41.
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 888–905.
Soltanisehat, L., Alizadeh, R., Hao, H., & Choo, K.-K.R. (2020). Technical, temporal, and spatial research challenges and opportunities in blockchain-based healthcare: A systematic literature review. IEEE Transactions on Engineering Management, 1–16.
Tang, C., Zhu, X., Liu, X., Li, M., Wang, P., Zhang, C., & Wang, L. (2019). Learning a joint affinity graph for multiview subspace clustering. IEEE Transactions on Multimedia, 21, 1724–1736.
Tian, F., Wang, Q., Li, X., & Sun, N. (2019). Heterogeneous multimedia cooperative annotation based on multimodal correlation learning. Visual Communication and Image Representation, 58, 544–553.
Toyama, K., Logan, R., & Roseway, A. (2003). Geographic location tags on digital images. ACM international conference on multimedia (pp. 156–166).
Valcarce, D., Parapar, J., & Barreiro, A. (2018). A mapreduce implementation of posterior probability clustering and relevance models for recommendation. Engineering Applications of Artificial Intelligence, 75, 114–124.
Verma, Y., & Jawahar, C.V. (2013). Exploring SVM for image annotation in presence of confusing labels. Proceedings British machine vision conference (p. 11747523).
Wang, D., Li, J., & Zhu, S. (2020). Detecting urban hot regions by using massive geo-tagged image data. *Neurocomputing* (pp. In Press, Corrected Proof).
Wang, H., Yang, Y., & Liu, B. (2019). GMC: Graph-based multi-view clustering. IEEE Transactions on Knowledge and Data Engineering, 32, 1116–1129.
Wang, R., Xie, Y., Xue, J.Y.L., & Zhang, M.H.Q. (2017). Large scale automatic image annotation based on convolutional neural network. Visual Communication and Image Representation, 49, 213–224.
Weinberger, K.Q., & Saul, L.K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10, 207–244.
Xing, E. P., Ng, A., Jordan, M. I., & Russell, S. J. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 521–528).
Xing, H., Meng, Y., Hou, D., Song, J., & Xu, H. (2017). Employing crowdsourced geographic information to classify land cover with spatial clustering and topic model. Remote Sensing, 9, 602.
Xue, Z., Li, G., & Huang, Q. (2018). Joint multi-view representation and image annotation via optimal predictive subspace learning. Information Sciences, 452, 180–194.
Yu, F., & Ip, H.H.S. (2006). Automatic semantic annotation of images using spatial hidden markov model. IEEE international conference on multimedia and expo (pp. 305–308).
Zamiri, M., & Yazdi, H.S. (2020). Image annotation based on multi-view robust spectral clustering. Visual Communication and Image Representation, 103003.
Zhang, C., Fu, H., Liu, S., Liu, G., & Cao, X. (2015). Low-rank tensor constrained multiview subspace clustering. IEEE international conference on computer vision (pp. 1582–1590).
Zhang, R., Zhang, L., Wang, X.-J., & Guan, L. (2011). Multi-feature pLSA for combining visual features in image annotation. International conference on multimedia (pp. 1513–1516).
Zhang, X., Zhao, Z., Zhang, H., Wang, S., & Li, Z. (2018). Unsupervised geographically discriminative feature learning for landmark tagging. Knowledge-Based Systems, 149, 143–154.
Zhao, M., Chow, W.S.T., Zhang, Z., & Li, B. (2015). Automatic image annotation via compact graph based semi-supervised learning. Knowledge-Based Systems, 76, 148–165.
Zhao, Y., Zhao, Y., & Zhu, Z. (2009). TSVM-HMM: Transductive SVM based hidden markov model for automatic image annotation. Expert Systems with Applications, 36, 9813–9818.
Zheng, L., Caiming, Z., & Caixian, C. (2018). MMDF-LDA: An improved multi-modal latent dirichlet allocation model for social image annotation. Expert Systems With Applications, 104, 168–184.
Zhou, N., Xu, Y., Cheng, H., Yuan, Z., & Chen, B. (2019). Maximum correntropy criterion based sparse subspace learning for unsupervised feature selection. IEEE Transactions on Circuits and Systems for Video Technology, 29, 404–417.