Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Image annotation based on multi-view robust spectral clustering☆

Mona Zamiri, Hadi Sadoghi Yazdi *

Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
Center of Excellence on Soft Computing and Intelligent Information Processing, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Article history:
Received 26 January 2019
Received in revised form 25 October 2020
Accepted 11 December 2020
Available online xxx

Keywords
Image annotation
Geo-tagged photos
Recommender systems
Maximum correntropy criterion
Multi-view spectral clustering
Geographical information

ABSTRACT

Nowadays, image annotation has been a hot topic in the semantic retrieval field due to the abundant growth of digital images. The purpose of these methods is to realize the content of images and assign appropriate keywords to them. Extensive efforts have been conducted in this field, which effectiveness is limited between low-level image features and high-level semantic concepts. In this paper, we propose a Multi-View Robust Spectral Clustering (MVRSC) method, which tries to model the relationship between semantic and multi-features of training images based on the Maximum Correntropy Criterion. A Half-Quadratic optimization framework is used to solve the objective function. According to the constructed model, a few tags are suggested based on a novel decision-level fusion distance. The stability condition and bound calculation of MVRSC are analyzed, as well. Experimental results on real-world Flickr and 500PX datasets, and Corel5K confirm the superiority of the proposed method over other competing models.

1. Introduction

The development of digital cameras and smartphones, as well as the popularity of social media and a great variety of them (e.g., Instagram and Flickr), have led to considerable growth of digital images. These digital image files may contain the camera or smartphone metadata, such as geographical coordinates, date of images' capture, and taken time. Geographical information (i.e., latitude and longitude) of images is one of the most popular metadata, which has attracted increased attention in recent years. Geo-tagged resources (e.g., video and image) have been widely used in many approaches [1–3]. Also, extensive work has been conducted on the influence of geo-information into tagging tasks [4–6]. Lee et al. [7] considered the relationship between the tag and geo-tag information and showed that there is a strong correlation between them. Toyama et al. [8] provided some reasons that images' content and location information have a real collaboration..

Due to the popularity and accessibility of these geo-tagged datasets, many methods have been proposed in this field (e.g., point-of-interest and tourism attractions recommendation [1,9–11], analyzing social metadata [12,13], geo-location prediction of input images [14,15], image annotation and tag recommendation [4–6,16]). Some researches are surveyed in the field of image annotation [17–19] and geo-tagged application [20,21].

Image annotation or tagging methods have been established to facilitate the management, retrieval, and searching of the images. These methods can be considered as a multi-label classification problem that tries to suggest suitable keywords to the input images. Generally, they consist of three phases: 1) feature extraction which expresses the content of images in the form of feature vectors, 2) learning phase which tries to model the relevance of images based on their extracted features, and 3) prediction phase that determines suitable tags for each test data via constructed model.

Multi-view feature fusion methods play a pivotal role in image annotation. Most prior Content-Based-Image-Retrieval (CBIR) methods try to extract multiple visual features to model the images. They usually suggested relevant tags that describe the content of images directly (Lei et al.[22] called these keywords as directly content-related tags), such as "*tree*" and "*flower*". These tagging models generally utilized low-level visual features, like color and texture, to suggest relevant tags. However, these features and textual tags are different naturally- semantic gap- that is to say, for example, they cannot differentiate between "*water*" and "*sky*" because of the similar color and texture. It is believed that textual tags are as significant as visual features in image tagging. For instance, as shown in Fig. 1, textual tags are "*sunset*", "*sky*", "*clouds*", "*building*", and "*washington*". There are common features between visual features and textual tags (i.e., "*sunset*", "*sky*", "*clouds*", "*building*"). Therefore, it is essential to model the correspondence between visual features and textual tags for image annotation. Moreover,

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
E-mail addresses: zamiri.mona@mail.um.ac.ir (M. Zamiri); h-sadoghi@um.ac.ir (H.S. Yazdi)

https://doi.org/10.1016/j.jvcir.2020.103003
1047-3203/

Fig. 1. An example image from Flickr photo-sharing website. Textual tags are useful for image annotation. "*sunset*", "*sky*", "*clouds*", and "*building*" are the common features of image and textual tags.

in real-world datasets, images can be described by indirectly content-related tags, for example, the geographical tag "*washington*" in Fig. 1. It is not easy to comprehend such tags from solely low-level visual features. Images in the same location generally follow the same patterns [8]. Thus, the correspondence between visual features of an image and its geospatial information should be employed to improve the effectiveness of image tagging for such datasets. Furthermore, a robust model should be proposed to improve the performance of image annotation because user-generated content is usually corrupted by noise. In addition, since supervised methods suffer from a vast manual labeling workload, they are not an appropriate choice for image annotation in practice. As a result, an unsupervised robust multi-view method should be used. Among different kinds of clustering methods, spectral clustering is of great interest due to its effective ability to handle nonlinearly separable datasets [23–25], and it is easy to implement. In summary, the contributions of this paper can be listed as follows:

- We present a new multi-view image annotation model based on feature-level fusion and Spectral Clustering (SC) in the training phase and decision-level fusion in the tag prediction phase.
- A novel multi-view robust similarity measure is proposed, which employs feature-level fusion to analyze the relations between training images and their semantic concepts. This similarity measure is utilized in the graph construction step of the SC.
- Geographical information of two image datasets is employed as useful indirectly content-related tags.
- An iterative algorithm is provided for constructing the similarity (affinity) matrix in the training phase based on the Maximum Correntropy Criterion (MCC) and half-quadratic (HQ) optimization method.
- Stability analysis and computational complexity of the proposed method are considered.
- Comprehensive experiments on two real-world datasets (Flickr and 500PX) and the Corel5K demonstrate the effectiveness of the proposed method over other state-of-the-art multi-view clustering models.

The rest of the paper is organized as follows. Section (2) reviews related works in image annotation, multi-view clustering methods, and correntropy-based models. Section (3) introduces the overall framework of the proposed method and some related discussions. Section (4) presents the experimental evaluation and discusses the most representative results. Finally, Section (5) concludes the paper.

2. Related work

In this section, a brief review of existing image annotation methods, graph-based multi-view clustering models, and the correntropy framework are considered separately.

2.1. Overview of image annotation models

In recent years, many approaches have been proposed in the task of image annotation. Generally, these approaches can be divided into three groups: 1) generative models, 2) discriminative methods, and 3) search-based algorithms.

The first group assumes that images have a specific distribution and tries to estimate its parameters. This group has two types of models: relevance and topic models. Relevance models generate the joint distribution of image features and tags. This group tries to compute the posterior probability of tags for untagged images and assign the tag with the highest probability [26,27]. Topic models try to explore topics in images and model images based on these topics [28,29]. Non-negative Matrix Factorization (NMF)-based methods fall into this category [30].

The second group learns a classifier for each tag. This classifier is learned based on different learning algorithms (e.g., Neural Networks (NN) [31,32] or Support Vector Machine (SVM) [33]).

The third group is search-based methods. These methods, like k-Nearest Neighbor (kNN), fall into the category of lazy learning methods, which means that they do not construct any model in their training phase. It determines the tags of a given query image based on the similarity between that image and training data. Makadia et al. [34] proposed the Joint Equal Contribution (JEC) model, which finds similar neighbors of the query image using low-level features and a combination of some distance measurements. Tags are retrieved from nearest neighbors or different aspects like co-occurrence factor. Verma et al. [35] provided a two-step variant of kNN (2PKNN) that considers image-to-tag and image-to-image similarities. They also proposed a metric learning framework that learns the weights of features.

Many approaches have been proposed as a hybrid model that utilizes the advantages of more than one group. Kalayeh et al. [36] proposed the NMF-KNN method to solve the problem of the continuous increase in data and tags. It has been specifically learned a model for each image. Guillaumin et al. [37] introduced a nearest-neighbor method, which is combined with metric learning by maximizing log-likelihood of tag prediction in the training data. Rad et al. [38] considered the problem of image annotation using the generative model and a search-based algorithm. This method extracts the latent factors and represents data to a low-rank latent factor space using NMF. It predicts tags using a search-based method in this space. They also proposed an annotation method, which is a hybrid model of generative and search-based methods [30]. For each view, their method finds a latent space by NMF and allows it to choose its number of basis factors. Finally, a weighted nearest neighbor based on a unified distance matrix is applied to predict the tags for the query images. Yang et al. [39] proposed the MVSAE method to build the correlations between low-level image features and high-level semantic concepts. Also, they provided a mechanism for solving the image keywords with imbalance distribution. Murthy et al. [40] introduced a hybrid method combining generative and discriminative models for image annotation application. It uses SVM and DMBRM to address the problems of irrelevant keywords and imbalanced data respectively. Zhao et al. [41] proposed a compact graph-based semi-supervised learning for image annotation. It identifies the neighborhood of each data and then reconstructs each sample based on its neighbors to learn a compact graph.

The performance of image annotation methods is very dependent on the images' feature descriptors; powerful descriptors can help to understand the images' content better and achieve more effective results.

Due to the importance of descriptors, the multi-view methods are provided to consider the content of images from different aspects [30,42,39,36]. In this way, useful features are extracted and fused based on different methods.

2.2. Overview of graph-based multi-view clustering models

With the advancement of technology, data are gathered from various sources or obtained via different feature descriptors. The complementary principle states that each view of features may contain useful knowledge that other views do not have. Thus, multiple views should be utilized so as to describe data samples accurately and comprehensively. Images shared on the websites, for example, have corresponding geographical and textual tags. Moreover, pictures can be described via different types of features. These data are well-known as multi-view data that contain various features with complementary information and have arisen in different fields, like data mining and pattern recognition. Many multi-view clustering methods have been proposed in the last few years. Multi-view graph clustering models seek to find a similarity fusion graph across available views and then apply different graph-cut algorithms (such as spectral clustering) on the similarity graph to produce the clustering results [43–54]. Nie et al. [43] proposed a graph-based multi-view learning method conducting clustering and local structure learning simultaneously. Cai et al. [44] proposed a graph-based multi-view spectral clustering method to integrate heterogeneous features. Also, Nie et al. [45] employed a parameter-free graph learning model, in which weight for each graph is learned automatically. Besides, a spectral rotation method for multi-view data is proposed in [46]. Wang et al. [48] proposed a novel graph-based learning method (GMC), which can construct the graph of each view and the unified graph matrix in a mutual reinforcement manner. Kang et al. [49] provided a multi-view spectral model, GFSC, which learns graph fusion and spectral clustering simultaneously. Tang et al. [50] deployed the low-rank representation (LRR) model to construct a joint similarity graph for multi-view subspace clustering. It used a diversity regularization term to learn the optimal weights of each feature view. Peng et al. [51] proposed a novel cross-view matching clustering (COMIC) without parameter selection, which can satisfy geometric consistency and cluster assignment consistency. Wen et al. [52] provided a novel incomplete multi-view clustering framework, GIMC-FLSD, based on matrix factorization to consider the local geometric and the unbalanced discriminating powers of incomplete views simultaneously. Liu et al. [53] provided an efficient and effective incomplete multi-view clustering method (EE-IMVC) and employed prior knowledge to regularize a learned consensus clustering matrix with linear computational complexity. Moreover, they proposed a late fusion incomplete multi-view clustering method [54] (LF-IMVC) which jointly learns a consensus clustering matrix, imputes incomplete matrices, and optimizes the corresponding permutation matrices within a three-step iterative algorithm.

Although previous multi-view clustering methods have achieved promising results, there are still some limitations. First, most of the prior graph-based clustering methods learn an individual similarity graph for each view and then fuse them to find the final clusters, like [44]. This approach fails to capture the relationship of data samples comprehensively since it ignores the complementary information across different views. Second, most current graph-based clustering methods construct the similarity matrix of each view separately and keep the constructed matrix fixed during fusion step [44,45,47]. Third, real-world datasets often contain random noises and outliers, which adversely deteriorate the similarity matrix and clustering performance. Most of the previous multi-view clustering methods usually employ non-robust distances (e.g., the Euclidean distance) to construct their similarity matrix [43,49]. Also, some methods assumed that the noise term has a sparse representation and used a sparse representation method for similarity graph construction [48,50].

To tackle the aforementioned limitations, we propose a novel multi-view spectral clustering based on the maximum correntropy criterion for image annotation, termed MVRSC, which could outperform other multi-view clustering methods even in real-world datasets. The maximum correntropy criterion (MCC) [55] has been widely applied to different adaptive algorithms with the purpose of mitigating the impulsive noises. Maier et al. [56] showed that the clustering performance highly relies on the quality and choice of the similarity graph. Therefore, a correntropy-based metric should be used for handling large outliers and noises. The proposed method constructs a robust similarity graph via an iterative MCC-based algorithm and the importance of each view is controlled by scaling factors (i.e., $\{\eta_v\}_{v=1}^{V}$) as well as automatically generated weights for each data sample of views (i.e., parameter "q").

2.3. Overview of correntropy-based models

Correntropy [55] is a concept that comes from information-theoretic learning (ITL), trying to deal with large outliers and has been widely employed to many areas, such as pattern recognition [57,58], computer vision [59,60] and signal processing [61,62]. The MCC-based models are robust against large outliers, and their solutions are more accurate than conventional Mean Square Error (MSE) ones. The correntropy measures the similarity of two arbitrary random variables X and Y. In practical applications, based on a finite number of samples $\{x_i, y_i\}_{i=1}^{n}$, this similarity measure is defined as follows:

$$\widehat{V}_\sigma(X,Y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma\left(x_i - y_i\right), \tag{1}$$

where $k_\sigma = g(X,\sigma) = \exp\left(-X^2/2\sigma^2\right)$ is the Gaussian kernel and σ is the kernel size. The maximum of Eq. (1) is called the maximum correntropy criterion (MCC) and has been used in different learning models. Zhou et al. [57] proposed a new robust principal component analysis (PCA) model based on maximum correntropy to eliminate the detrimental effects of outliers. A novel robust subspace learning model in an unsupervised scenario is also presented in [59] for feature selection based on the MCC to combat the negative effects of the noises, which are produced by the sensors.

The proposed method is a hybrid model that employs the discriminative method and search-based method together. We use multiple features (geo-information of images is one of them) in the form of multiple matrices and provide an optimization problem based on the MCC to construct the similarity matrix and learn the weights of each view iteratively. Then, we determine the clusters and representative tags for each cluster. Finally, for each test image, tags are suggested using a search-based method and a new decision-level fusion.

3. Proposed method: multi-view robust spectral clustering

In this section, we introduce the proposed multi-view robust spectral clustering (MVRSC). In general, fusion techniques can be divided into two categories: feature-level fusion and decision-level fusion [63].

On the one hand, the feature-level fusion method combines different extracted features (e.g., visual, textual, geographical features) into a single feature vector and analyzes this combined feature vector. The advantage of this fusion approach is that the aggregation of different features at an early step can provide more suitable results since each view can co-regularize with other views during the clustering algorithm.

On the other hand, in the decision-level fusion methods, the extracted features from different sources are classified independently, and the results are combined as a final decision vector. The advantage of such an approach is the simple comparison of the decisions obtained

from various sources. The proposed method uses feature-level and decision-level fusion methods so as to employ the advantages of both methods.

As shown in Fig. 2, the overall structure of our model consists of three steps: 1) feature extraction, which contains different features obtained by different feature extraction techniques, 2) robust modeling based on feature-level fusion, and 3) tag prediction based on decision-level fusion. The first step describes the content of images using different descriptors and their metadata. The second step is the basic HQ-spectral clustering, which is used to construct the mapping from images' extracted features to their tags. In this phase, we use a novel feature-level fusion technique that unifies different extracted features. In the final step, a few tags are suggested for the unknown test data based on the constructed model. It uses a decision-level fusion approach to obtain the final decision. In the following, these steps will be explained in more detail.

3.1. Robust modeling based on feature-level fusion

First, let us introduce some notational conventions which are used throughout the paper. Matrices are shown in boldface capital letters (e.g., $\mathbf{X}$). Vectors and scalars are denoted in boldface lowercase letters (e.g., $\mathbf{x}$) and lowercase letters (e.g., x) respectively. Given an arbitrary matrix $\mathbf{X} \in R^{m\times n}, x_{i,j}$ and $\mathbf{x}_i$ represent its (i,j)-th entry and i^{th} column respectively. As shown in Table 1, let N denotes the number of training images. $\{\mathbf{F_v}\}_{v=1}^{V} \in R^{\dim(v)\times N}$ is the data matrix of the v^{th} view, $\dim(v)$ is its feature dimension. Let define the similarity matrix $\mathbf{C}$ as follows:

$$\mathbf{C} = \begin{bmatrix} 0 & c_{12} & c_{13} & \dots & c_{1N} \\ c_{21} & 0 & c_{23} & \dots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{N1} & c_{N2} & c_{N3} & \dots & 0 \end{bmatrix} \in R^{N\times N},$$

$\mathbf{c_i} \in R^{(N-1)\times 1}$ represents the i^{th} column vector of matrix $\mathbf{C}$ (except the diagonal element) which is obtained based on the similarity between the i^{th} training image and other training samples. Now, let $\{\mathbf{F_v^i}\}_{v=1}^{V}$ corresponds to different matrices for the i^{th} sample based on different views, defined as follows:

$$\mathbf{F_v^i} = \begin{bmatrix} ndist\left(\mathbf{f_v^1}, \mathbf{f_v^i}\right) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & ndist\left(\mathbf{f_v^N}, \mathbf{f_v^i}\right) \end{bmatrix} \in R^{(N-1)\times(N-1)},$$

where $\mathbf{f_v^k}, k = \{1, \dots, N\}$ represents the k^{th} column of matrix $\mathbf{F_v}$. $\mathbf{F_v^i}$ is a diagonal matrix having normalized distances between the i^{th} sample and other images (except the i^{th} data) based on the v^{th} view on the diagonal. To prevent scale difference caused by different views, we normalize the $\{\mathbf{F_v}\}_{v=1}^{V}$ data matrices to have a unit of l_2-norm (e.g., $ndist(\mathbf{a}, \mathbf{b})$ calculates l_2-norm of two normalized vectors $\mathbf{a}$ and $\mathbf{b}$).

For each data i and view v, an error is defined as:

$$e_{vi} = \left\|\mathbf{F_v^i c_i}\right\|_2, v = \{1, 2, \dots, V\}, i = \{1, \dots, N\}, \tag{2}$$

where $\|.\|_2$ is the l_2-norm of a vector and e_{vi} states the self-expressiveness property, in which each sample can be reconstructed through a linear combination of other samples and the combination coefficients (i.e., $\{\mathbf{c_i}\}_{i=1}^{N}$) represent the similarity between samples [64,65]. According to the Eq. (1), we define the correntropy of the error function for v^{th} view as follows:

$$g(e_v) = \frac{1}{N}\sum_{i=1}^{N} \exp\left(-\frac{1}{2}\eta_v e_{vi}^2\right), \quad v = \{1, 2, \dots, V\}, \tag{3}$$

where η_v is the scaling factor for v^{th} view.

3.1.1. MCC-optimization problem

We define the below maximum correntropy cost function based on Eq. (3):

$$\max_{\mathbf{c_i}} J(\mathbf{c_i}) = \max_{\mathbf{c_i}} \sum_{i=1}^{N}\sum_{v=1}^{V} \exp\left(-\frac{1}{2}\eta_v e_{vi}^2\right) \quad s.t. \begin{cases} \mathbf{1}^T\mathbf{c_i} = 1 \\ \mathbf{c_i} \geqslant 0, \quad i = 1, \dots, N \end{cases}. \tag{4}$$

where $\mathbf{1} \in R^{(N-1)\times 1}$ is a vector and all of its elements are 1. HQ modeling is described via the conjugate function theory [66,67]. So, we

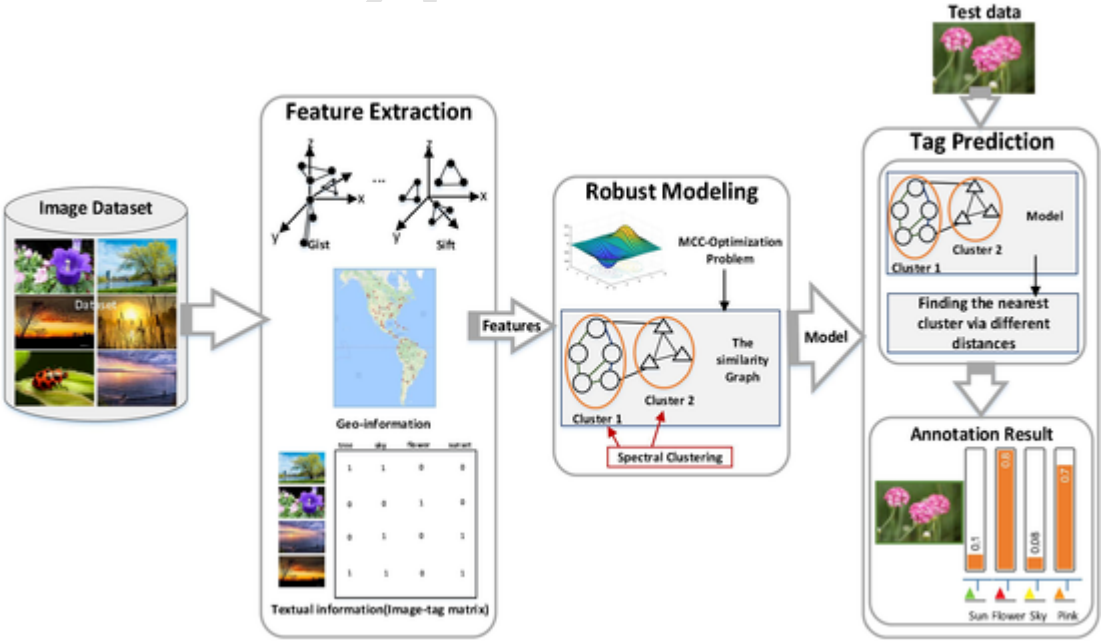


Fig. 2. The structure of our image annotation method. The proposed method has three phases: the first phase is feature extraction, the second phase constructs the model based on training data, and in the third phase, using the constructed model, some tags are predicted for each test data.

Table 1
Notation.

Symbol	description
N	Number of training images
$\mathbf{C}$	Similarity/Affinity matrix
V	Number of accessible views v = {1,…,V}
$\mathbf{F_v^i}$	Matrix related to the v^{th} views of the i^{th} image
$ndist(\mathbf{a},\mathbf{b})$	Normalized distance between vectors $\mathbf{a}$ and $\mathbf{b}$
t	Number of tags

solve the cost function $J(\mathbf{c_i})$ in Eq. (4) using the half-quadratic problem solving method, as follows:

Proposition 1: A convex conjugate function ϕ of $g(e)$ exists such that (see Appendix A):

$$\begin{aligned} g(e) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{v=1}^{V}\sup_{p_{vi}<0}\left(\frac{1}{2}\eta_v e_{vi}^2 p_{vi} - \phi(p_{vi})\right), \\ &= \frac{1}{N}\sup_{p_{vi}<\mathbf{0}}\left\{\sum_{i=1}^{N}\sum_{v=1}^{V}\left(\frac{1}{2}\eta_v e_{vi}^2 p_{vi} - \phi(p_{vi})\right)\right\}, \end{aligned} \tag{5}$$

where convex function $\{\phi(p_{vi})\}_{v=1}^{V}$ is defined as $\phi(p_{vi}) = -p_{vi}\log(-p_{vi}) + p_{vi}$. The second equation in Eq. (5) establishes since $\frac{1}{2}\eta_{vi}^2 e_{vi}^2 p_{vi} - \phi(p_{vi}), i = 1,\dots,N, v = 1,\dots,V$ are independent functions in terms of p_{vi}. Thus, Eq. (4) is equivalent to:

$$\begin{aligned} \max_{\mathbf{c_i}, \mathbf{p_v}} J(\mathbf{c_i}, \mathbf{p_v}) &= \max_{\mathbf{c_i}, \mathbf{p_v}} \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\eta_v e_{vi}^2 p_{vi} - \phi(p_{vi}) \\ s.t. &\begin{cases} \mathbf{1}^T\mathbf{c_i} = 1 \\ \mathbf{c_i} \geqslant 0 \quad i = 1,\dots,N \end{cases}. \end{aligned} \tag{6}$$

When $\mathbf{c_i}$ is fixed, the following equation holds:

$$\max_{\mathbf{c_i}} J(\mathbf{c_i}) = \max_{\mathbf{c_i}, \mathbf{p_1}, \mathbf{p_2}, \dots, \mathbf{p_V}} \widehat{J}(\mathbf{c_i}, \mathbf{p_1}, \mathbf{p_2}, \dots, \mathbf{p_V}). \tag{7}$$

We can solve the above equation using the alternating optimization method. In other words, given $\mathbf{c_i}$ we optimize over p_{vi} and then given p_{vi}, we solve the optimization problem over $\mathbf{c_i}$ in an iterative manner. Thus, we have:

$$\begin{aligned} &\mathbf{step1}: \max_{\mathbf{p_1}, \mathbf{p_2}, \dots, \mathbf{p_V}} \widehat{J}(\mathbf{c_i}, \mathbf{p_1}, \dots, \mathbf{p_V}) \\ \rightarrow p_{ji} &= -\exp\left(-\frac{1}{2}\eta_j e_{ji}^2\right), \quad j = 1,\dots,V, \quad i = 1,\dots,N. \\ &\mathbf{step2}: \max_{\mathbf{c_i}} \widehat{J}(\mathbf{c_i}, \mathbf{p_1}, \dots, \mathbf{p_V}) \\ \rightarrow \max_{\mathbf{c_i}} J(\mathbf{c_i}) &= \max_{\mathbf{c_i}} \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\eta_v e_{vi}^2 p_{vi}. \end{aligned} \tag{8}$$

By defining $\begin{aligned} q_{ji} &= -p_{ji} \\ &= \exp\left(-\frac{1}{2}\eta_j e_{ji}^2\right), \quad j = 1,\dots,V, \quad i = 1,\dots,N \end{aligned}$:

$$\min_{\mathbf{c_i}} J(\mathbf{c_i}) = \min_{\mathbf{c_i}} \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\eta_v e_{vi}^2 q_{vi}. \tag{9}$$

Since $\eta_1,\dots,\eta_V$ are constant, by expanding Eq. (9) we have:

$$J(\mathbf{c_i}) = \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\left\|\mathbf{F_v^i}\mathbf{c_i}\right\|_2^2 q_{vi}. \tag{10}$$

Finally, Eq. (10) is converted to the below equation:

$$\begin{aligned} \min_{\mathbf{c_i}} J(\mathbf{c_i}) &= \min_{\mathbf{c_i}} \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\left(\mathbf{c_i}^T\left(\mathbf{F_v^{i}}^T\mathbf{F_v^i}q_{vi}\right)\mathbf{c_i}\right) \\ s.t. &\begin{cases} \mathbf{1}^T\mathbf{c_i} = 1 \\ \mathbf{c_i} \geqslant 0, \quad i = 1,\dots,N \end{cases} \end{aligned} \tag{11}$$

Eq. (11) can be rewritten in the following quadratic programming (QP) problem:

$$\begin{aligned} \min_{\mathbf{c_i}} J(\mathbf{c_i}) &= \min_{\mathbf{c_i}} \sum_{i=1}^{N}\sum_{v=1}^{V}\frac{1}{2}\left(\mathbf{c_i}^T\mathbf{H_{iv}}\mathbf{c_i}\right) + \mathbf{f}^T\mathbf{c_i} \\ s.t. &\begin{cases} \mathbf{a_{eq}}\mathbf{c_i} = b_{eq} \\ \mathbf{A}\mathbf{c_i} \leqslant \mathbf{b}, \quad i = 1,\dots,N \end{cases}, \end{aligned} \tag{12}$$

which its parameters are:

- $\mathbf{A}$ is a $(N-1)\times(N-1)$ diagonal matrix that diagonal elements are −1.
- $\mathbf{H_{iv}} = \left(\mathbf{F_v^{i}}^T\mathbf{F_v^i}q_{vi}\right) \in R^{(N-1)\times(N-1)}$.
- $\mathbf{b} = \left[0, 0, \dots, 0_{|N-1|}\right]^T \in R^{(N-1)}$.
- $\mathbf{f} = \mathbf{b}, \mathbf{a_{eq}} = \left[1, \dots, 1_{|N-1|}\right] \in R^{1\times(N-1)}$ is a full-row-rank matrix, $b_{eq} = 1$.

Eq. (12) can be converted into:

$$\begin{aligned} \min_{c_{i,j}} J(c_{i,j}) &= \min_{c_{i,j}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{v=1}^{V}\left(c_{ij}{}^2 q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2\right) \\ s.t. &\begin{cases} \sum_{j=1}^{N} c_{ij} = 1, \quad \forall i = 1,\dots,N \\ 0 \leqslant c_{ij} \leqslant 1 \end{cases}. \end{aligned} \tag{13}$$

The Lagrangian of the above equation takes the form:

$$\begin{aligned} L(\mathbf{C}, \lambda) = &\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{v=1}^{V}\left(c_{ij}{}^2 q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2\right) \\ &-\sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{N} c_{ij} - 1\right). \end{aligned} \tag{14}$$

The optimal c_{ij} can be solved by setting $\frac{\partial L(\mathbf{C},\lambda)}{\partial c_{ij}} = 0$, which yields:

$$\begin{aligned} \frac{\partial L(\mathbf{C},\lambda)}{\partial c_{ij}} = 0 &\rightarrow \frac{1}{2}\sum_{v=1}^{V} 2 c_{ij} q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2 - \lambda_i = 0 \\ &\rightarrow c_{ij}\sum_{v=1}^{V} q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2 = \lambda_i \\ &\rightarrow c_{ij} = \frac{\lambda_i}{\sum_{v=1}^{V} q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2}. \end{aligned} \tag{15}$$

With replacement of the Eq. (15) into the constraint $\sum_{j=1}^{N} c_{ij} = 1$:

$$\lambda_i = \frac{1}{\sum_{j=1}^{N}\frac{1}{\sum_{v=1}^{V} q_{vi}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2}}, \tag{16}$$

$$c_{ij} = \frac{1}{\sum_{m=1}^{N}\frac{\sum_{v=1}^{V}\exp\left(-\frac{1}{2}\eta_v e_{vi}^2\right)\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2}{\sum_{v=1}^{V}\exp\left(-\frac{1}{2}\eta_v e_{vi}^2\right)\left(ndist\left(\mathbf{f_v^m}, \mathbf{f_v^i}\right)\right)^2}}. \tag{17}$$

Note that e_{vi} is a function of the error, which depends on the coefficient vector $\mathbf{c_i}$ ($e_{vi} = \left\|\mathbf{F_v^i}\mathbf{c_i}\right\|_2$), so Eq. (17) is an iterative equation. It can be expressed as $c_{k+1} = h(c_k)$, where k denotes the iteration number and $h(c_k)$ is:

$$\begin{aligned} h(c_k) &= \frac{1}{\sum_{m=1}^{N}\frac{\sum_{v=1}^{V}\alpha_{i,v}\beta_{i,j,v}}{\sum_{v=1}^{V}\alpha_{i,v}\gamma_{i,m,v}}}, \\ &\begin{cases} \alpha_{i,v} = \exp\left(-\frac{1}{2}\eta_v\sum_{l=1}^{N}\left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)(c_{li})_k\right)^2\right), \\ \beta_{i,j,v} = \left(ndist\left(\mathbf{f_v^j}, \mathbf{f_v^i}\right)\right)^2, \\ \gamma_{i,m,v} = \left(ndist\left(\mathbf{f_v^m}, \mathbf{f_v^i}\right)\right)^2. \end{cases} \end{aligned} \tag{18}$$

An approach to finding the optimal solution is the following equation:

$$(c_{ij})_{k+1} = h\left((c_{li})_k\right), \tag{19}$$

where $(c_{li})_k$ denotes the coefficient based on the similarity between the l^{th} and the i^{th} sample at iteration k.

According to the Taylor series expansion and its first degree approximation ($\exp(-x) = 1 - x$), the fixed-point algorithm Eq. (19) becomes:

$$\begin{cases} \left(c_{ij}\right)_{k+1} = \widehat{h}\left(\left(c_{li}\right)_k\right), \\ \widehat{h}\left(\left(c_{li}\right)_k\right) = \frac{1}{1+\sum_{m=1,m\neq j}^{N}\frac{\sum_{v=1}^{V}\widehat{\alpha_{i,v}}\beta_{i,j,v}}{\sum_{v=1}^{V}\widehat{\alpha_{i,v}}\gamma_{i,m,v}}}, \\ \widehat{\alpha_{i,v}} = 1-\frac{1}{2}\eta_v\sum_{l=1}^{N}\left(ndist\left(\mathbf{f_v^l},\mathbf{f_v^i}\right)\left(c_{li}\right)_k\right)^2, \\ \beta_{i,j,v} = \left(ndist\left(\mathbf{f_v^j},\mathbf{f_v^i}\right)\right)^2, \\ \gamma_{i,m,v} = \left(ndist\left(\mathbf{f_v^m},\mathbf{f_v^i}\right)\right)^2. \end{cases} \tag{20}$$

Algorithm 1 Constructing Similarity Matrix **C**

Require: N training images are expressed by different $v = \{1, 2, \ldots, V\}$ views, $\{\eta_v\}_{v=1}^{V}$
Ensure: Similarity matrix **C**
Initialize: $q_{ji} = 1, \quad j = 1, \ldots, V, \quad i = 1, \ldots, N$.
for $k = 1$ to *Iter* **do**
 for $i = 1$ to N **do**
 Construct $\mathbf{F_v^i}, v = 1, \ldots, V$ matrices.
 Compute Quadratic programming problem in Eq. (12) and determine $\mathbf{c_i}$.
 Update all $q_{ji} = -p_{ji} = \exp\left(-\frac{1}{2}\eta_j e_{ji}^2\right), \quad j = 1, \ldots, V$.
 end for
end for

3.1.2. Unified multi-view feature for image representation

By solving the optimization problem of Eq. (12) and determining the matrix **C**, we represent training images in a new space that considers all views. In other words, the matrix **C** indicates the similarity between training samples based on multi-view feature fusion. The construction steps of the proposed similarity matrix are summarized in *Algorithm 1*.

Algorithm 2 Image Annotation Based on Multi-view Robust Spectral Clustering

Require: N training images are expressed by $v = \{1, 2, \ldots, V\}$ views' matrices, "t" and "K" are the numbers of annotated tags and clusters respectively.
Ensure: "t" candidate tags for annotating of each test image.
- Construct similarity matrix **C** according to the Algorithm 1.
- Apply spectral clustering on the constructed similarity matrix **C** [68] and extract representative concepts of the clusters.
- Compute the distance of the test data and clusters' centers according to the above-defined distances in Section 3.2.
- Assign "t" representative tags of the closest cluster to the test data.

3.1.3. Concept extraction using spectral clustering

Spectral clustering is a powerful unsupervised method that determines the clusters based on the partitioning of a similarity graph [68,23,69]. A spectral clustering algorithm generally consists of three steps: pre-processing, decomposition, and grouping. A similarity graph and Laplacian matrix are constructed for the dataset in the pre-processing step. In the decomposition step, the dataset is represented using the eigenvectors corresponding to the K (number of clusters) smallest eigenvalues of the Laplacian matrix. Finally, clusters are obtained from the new representation in the grouping step. There are different ways for similarity graph construction (e.g., k-nearest neighbor, ϵ-neighborhood and fully-connected graph). According to the Algorithm 1, our similarity graph (i.e., matrix **C**) is a fully-connected and also a directed (asymmetric) graph (i.e., $c_{i,j} \neq c_{j,i}$). In order to convert it to an undirected graph, we use the following equation [70]:

$$\mathbf{C} = \max\left(\mathbf{C}, \mathbf{C}^T\right). \tag{21}$$

We represent training images in the form of an undirected weighted graph $\mathbf{C} = \{Node, Edge\}$, where the nodes of the graph are training data, and the edges are weighted between each pair of nodes based on a similarity function. In the proposed method, the edges are weighted by the Eq. (12). To partition this graph, we use normalized-cut, which can formulate as generalized eigenvalue decomposition [69]. Using spectral clustering and partitioning data into "K" clusters, we extract concepts by ranking the most frequent words that are repeated in training data of each cluster. The number of clusters "K" is adjusted as follows [5]:

$$K = \sqrt{\frac{N}{2}}. \tag{22}$$

3.2. Tag prediction based on decision-level fusion

The last phase of the proposed method is tag prediction, which uses the constructed model to suggest suitable tags for unlabeled test images. In this phase, relevant tags are predicted based on the closest cluster's representative concepts for each test data. We assume that all V views except the textual view are available for all test images, and we aim to predict their textual tags based on different views.

By using different available views, different distances are defined based on the Euclidean distance, and for each one, the closest cluster is found. For example, when our available views for test data are geographical and visual, we list the below distances and for each distance, suggest different tags:

- **Visual distance**: In this case, the visual feature of a test image is compared with the visual centers of clusters, and the most similar cluster is determined.
- **Geographical distance**: In this method, the geographical feature of the test image is compared with the geographical centers of clusters, and the closest one is determined.
- **Geo-visual distance**: By concatenating the geographical and visual features, the geo-visual feature vector of the test image is compared with the geo-visual centers of clusters. Finally, it selects the closest cluster.
- **Fusion distance**: This method determines the nearest cluster by combining three above-mentioned distances and majority vote.

The proposed image annotation method is summarized in *Algorithm 2*.

3.3. Stability analysis and bound calculation of MVRSC

According to the Eq. (20), we have:

$$\widehat{h}\left(\left(c_{li}\right)_k\right) = \frac{1}{1+\sum_{m=1,m\neq j}^{N}\frac{\sum_{v=1}^{V}\widehat{\alpha_{i,v}}\beta_{i,j,v}}{\sum_{v=1}^{V}\widehat{\alpha_{i,v}}\gamma_{i,m,v}}}, \quad \begin{cases} \widehat{\alpha_{i,v}} = 1-\frac{1}{2}\sum_{l=1}^{N}\zeta_{l,v,i}\left(c_{li}\right)_k^2, \\ \zeta_{l,v,i} = \left(ndist\left(\mathbf{f_v^l},\mathbf{f_v^i}\right)\right)^2, \\ \beta_{i,j,v} = \left(ndist\left(\mathbf{f_v^j},\mathbf{f_v^i}\right)\right)^2, \\ \gamma_{i,m,v} = \left(ndist\left(\mathbf{f_v^m},\mathbf{f_v^i}\right)\right)^2. \end{cases} \tag{23}$$

Eq. (23) can be derived concisely, we introduce some symbols:

$$\widehat{h}\left(\left(c_{li}\right)_k\right) = \frac{S}{M}, \quad \begin{cases} S = \sum_{m=1,m\neq j}^{N}\sum_{v=1}^{V}\left[\gamma_{i,m,v} - \frac{1}{2}\gamma_{i,m,v}\sum_{l=1}^{N}\zeta_{l,v,i}\left(\left(c_{li}\right)_k\right)^2\right], \\ M = S + \left[(N-1)\sum_{v=1}^{V}\left(\beta_{i,j,v} - \frac{1}{2}\beta_{i,j,v}\sum_{l=1}^{N}\zeta_{l,v,i}\left(\left(c_{li}\right)_k\right)^2\right)\right]. \end{cases} \tag{24}$$

Theorem 1 *Stability Condition: MVRSC is stable if* $\exists\gamma_{i,m,v} \neq 0, \quad \forall i,m,v$ *or* $\exists\beta_{i,j,v} \neq 0, \quad \forall i,j,v$.

Proof Since the term $\sum_{l=1}^{N}\zeta_{l,v,i}\left(\left(c_{li}\right)_k\right)^2$ is always larger than or equal to zero; Consequently, non-zero denominator is the condition of bounded $\left(c_{li}\right)_k$. □

According to Eq. (24) if $\left(c_{li}\right)_k = 0, \quad \forall l, i = 1, \ldots, N$ then:

$$\left(c_{i,j}\right)_{k+1} = \widehat{h}(0),$$
$$\widehat{h}(0) = \frac{\sum_{m=1,m\neq j}^{N}\sum_{v=1}^{V}\gamma_{i,m,v}}{\sum_{m=1,m\neq j}^{N}\sum_{v=1}^{V}\gamma_{i,m,v}+(N-1)\sum_{v=1}^{V}\beta_{i,j,v}}. \tag{25}$$

Hence, according to Eq. (25) and definition of $\gamma_{i,m,v}$ and $\beta_{i,j,v}$ in Eq. (23), the denominator of Eq. (23) is always larger than zero and MVRSC is stable. This completes the proof. In order to calculate the bound of $\left(c_{li}\right)_k$, we take the limit of Eq. (24) in infinity:

$$\lim_{\left(c_{li}\right)_k\to\infty}\frac{S}{M} = \frac{\sum_{m=1,m\neq j}^{N}\sum_{v=1}^{V}\gamma_{i,m,v}\sum_{l=1}^{N}\zeta_{l,v,i}}{\sum_{m=1,m\neq j}^{N}\sum_{v=1}^{V}\gamma_{i,m,v}\sum_{l=1}^{N}\zeta_{l,v,i}+(N-1)\sum_{v=1}^{V}\beta_{i,j,v}\sum_{l=1}^{N}\zeta_{l,v,i}}. \tag{26}$$

3.4. Computational complexity of MVRSC

The main computation of Algorithm 2 is due to the construction of the similarity matrix (Algorithm 1) and calculation of eigenvalue decomposition of the corresponding Laplacian matrix. Specifically, constructing the similarity matrix costs about $O\left(N^2kV\right)$, where k is the number of iteration steps. The complexity of calculating the eigenvalue decomposition of the Laplacian matrix is $O\left(N^3\right)$ because of using the SVD method.

4. Experimental results

In this section, we implement different experiments on three image datasets to evaluate the quality of the proposed method (MVRSC). Firstly, we describe the datasets used in the experiments. Then, we introduce compared methods and evaluation metrics. Regarding the compared methods, we utilize the code released by the corresponding authors. In the experimental setting, we illustrate how features are extracted, and parameters are set for analysis. Finally, we demonstrate the effectiveness of the proposed method in the comparison results and analysis section.

4.1. Datasets

We use three image datasets (Flickr photos, 500PX, and Corel5k) to evaluate the proposed method. we explain each dataset in more detail, as follows:

1) **Flickr photos**: We collected 7,000 images with their metadata (i.e., textual tags and geographical coordinates) from Flickr[1] via its API [13] which is developed by the students of Pattern Recognition Laboratory of Ferdowsi University of Mashhad (FUM). Each image has six tags on average.
2) **500PX**: This real dataset is gathered from 500px website[2] through its API and is arranged by students of Pattern Recognition Laboratory of FUM. It contains 7,000 annotated images with their metadata (i.e., textual tags and geographical coordinates). Every image is annotated with eight tags on average.

Images were uploaded publicly by users of Flickr and 500PX websites. Fig. 3 shows the areas where these collected images are taken.

3) **Corel5k**: It consists of 5,000 images collected from the Corel CD set [71]. This dataset has been divided into 4,500 samples as a training set and 500 images for testing. Each image is annotated with 3.5 tags on average.

Example images of Flickr photos, 500PX, and Corel5K are shown in Fig. 4(a)-(c) respectively. Table 2 illustrates the datasets' information.

4.2. Compared methods and evaluation criteria

In order to evaluate the performance of our proposed method, standard evaluation metrics in information retrieval (i.e., Precision (*PR*), Recall(*RC*), and F-score ($F1$) [72]) are adopted for each tag. Finally, their reported results are averaged across all tags.

We implement different experiments to evaluate the proposed annotation method on Flickr and 500PX datasets. In this regard, we use different views (i.e., geographical, textual, and visual) and implement different modalities to evaluate the impact of each view in the proposed image annotation method. These modalities can be listed as follows:

- **G**: We run the proposed method by just using geographical features and annotate images based on their locations (v = {Geographical}).
- **V**: In this case, the proposed method is run by employing only visual features of images (v = {Visual}).
- **GV**: This modality tries to suggest proper tags to the images by fusing geographical and visual features of images in the training and prediction phase (v = {Geographical, Visual}).
- **VT**: In this case, the proposed method employs visual and textual features of training images for constructing the similarity graph (v = {Visual, Textual}) and then predicts tags based on the aforementioned visual distance.
- **GT**: The similarity graph of this modality is constructed via geographical and textual features of training images (v = {Geographical, Textual}). In the prediction phase, relevant tags of test images are recommended based on the aforementioned geographical distance.
- **MF**: This modality normalizes all available views and then builds a matrix via merging them with equal weight to construct the similarity graph. Then, the above-mentioned fusion distance is applied to predict tags for a new test image.
- **GTV**: This modality is the proposed method (MVRSC), which constructs the similarity graph by integrating all views (v = {Geographical, Textual, Visual}) in the training phase. Then, suitable tags are suggested to the test images via the aforementioned fusion distance.

Note that we cannot employ only textual features of images since, in this modality (i.e., v = {Textual}), there are no other views for annotating test images in the prediction phase. Therefore, we omit this modality from the above list.

To evaluate the effectiveness of the proposed method, we compare it with several multi-view image annotation tasks on the Corel5K dataset.

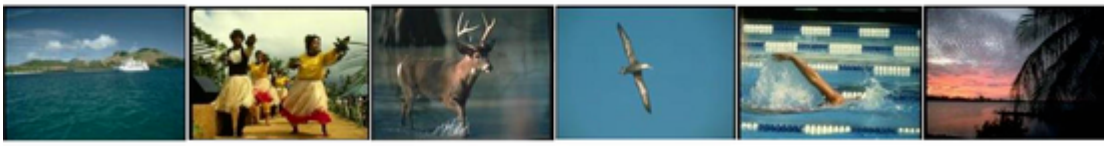
- NMF-KNN [36]: It is a hybrid model that combines the nearest neighbor method with multi-view non-negative matrix factorization.
- OPSL [42]: It is an optimal predictive subspace learning method which conducts multi-view representation and image annotation.
- MVSAE [39]: It uses a multi-view stacked auto-encoder framework to build the correlations between low-level image features and high-level semantic concepts using the sigmoid function predictor and iteration algorithm.
- MvNMF [30]: It uses multi-view graph regularization NMF with a different number of basis vectors for each view.

[1] http://www.flickr.com
[2] http://www.500px.com

Fig. 3. Our case study of geographical area. Locations of Flickr and 500PX images are depicted in the left and the right maps respectively.

(a) Flickr dataset sample images

(b) 500PX dataset sample images



(c) Corel5K dataset sample images

Fig. 4. Example images of Flickr, 500PX and Corel5K datasets are shown in a–c respectively.

Table 2
Information about datasets (#Feature).

#View	Flickr photos	500PX	Corel5k
1	Geographical (2)	Geographical (2)	DenseSift (1000)
2	Visual (4096)	Visual (4096)	GIST (512)
3	Textual (150)	Textual (117)	HSV (4096)
#Total image	7000	7000	5000
#Training image	5600	5600	4500
#Test image	1400	1400	500
#Tags per image (on average)	6	8	3.5
#Tags	150	117	260

We also compare the performance of the proposed method with both single view and multi-view clustering methods as follows:

- Spectral clustering (**SC**) [73]: We use the standard SC method as a baseline, and apply it to each view. SC(1) denotes the implementation of SC on the 1^{st} view.
- Multi-modal spectral clustering (**MMSC**)[44]: A multi-modal spectral clustering is provided in order to integrate heterogeneous image features and add a non-negative constraint to their objective function to improve the efficiency of clustering.
- Auto-weighted multiple graph learning (**AMGL**) [45]: A novel parameter-free framework is proposed that learns the similarity graphs and weights for different views automatically.

- Multi-view learning with adaptive neighbors (**MLAN**) [43]: In this approach, Multi-view clustering/semi-supervised classification and local structure learning are performed simultaneously. The weight of each view is also learned automatically.
- Multi-view clustering via adaptively weighted procrustes (**AWP**) [46]: It proposed an optimization strategy to provide an adaptively weighted procrustes method.
- Graph-based multi-view clustering (**GMC**) [48]: This method constructs the graph of each view and the fusion graph jointly to help each other in a mutual reinforcement manner.
- Multi-graph fusion for multi-view spectral clustering (**GFSC**) [49]: It proposed a model which performs graph fusion and spectral clustering simultaneously.
- Learning a Joint Affinity Graph for Multi-view Subspace Clustering [50]: It deployed the low-rank representation (LRR) method so as to learn a joint similarity graph of various views for multi-view subspace clustering.
- Cross-view matching clustering (**COMIC**) [51]: This method can automatically learn all parameters in a data-driven way.

4.3. Experimental setting

Annotation and tagging methods can be considered as multi-label problems. Therefore, using only one type of feature does not achieve a suitable result. For this purpose, different combinations of features representing data from different views (e.g., visual, geographical, and textual) are often used. Also, [42] showed that the multi-view annotation task achieves better performance than a single-view because of the widespread exploitation of different views. For Flickr photos and 500PX datasets, we employ three types of views: visual view, which is extracted with the AlexNet model [74] that is trained on the ImageNet dataset [75]; Geographical view (i.e., latitude and longitude) are available as their metadata. Textual view is extracted by the Term Frequency-Inverse Document Frequency (TF-IDF) method [76]. For the Corel5K dataset, we use three different types of visual features, which are local and global descriptors: SIFT [77], GIST, and HSV color spaces. We also utilize an annotation matrix as their textual features.

We compare the proposed method with different state-of-the-art methods and implement them by their parameter tuning steps and the experimental setting in their papers to make the results fair enough. In order to determine the proper parameters of each view (i.e., $\eta_1, \cdots, \eta_V$) in the proposed MVRSC method, we perform a grid search method on a random subset of training images for training and evaluation. We find the optimal value of parameters that yields an optimal model, which minimizes the average error of all views in all images. This error is defined as follows:

$$Average\_Error = \frac{1}{N * |v|} \sum_{j=1}^{V} \sum_{i=1}^{N} e_{ji}, \tag{27}$$

where $|v|$ denotes the number of available views.

4.4. Comparison results and analysis

The proposed method combines different views to retrieve relevant tags. To evaluate the performance of the proposed method on the Flickr and 500PX datasets, we evaluate it on each modality introduced earlier, individually. Moreover, we compare the proposed MVRSC model with different graph-based spectral clustering methods. Then, we carry out the experiments on the Corel5K dataset and compare it with several state-of-the-art methods.

Table 3 represents Precision, Recall, and F1 results for seven different modalities on the Flickr photos and 500PX datasets. This table shows that the integration of different views (i.e., GTV) can achieve a significant performance over other modalities on both datasets. Also, it is evident that the GV model achieves better performance compared to the other multi-view modalities (i.e., GT, VT, and MF); Because this model utilizes the feature-level and decision-level fusion based on available geographical and visual views in the training and prediction phase while other multi-view models use only one of these fusion techniques whether in the training or prediction phase.

Table 3
Recall, Precision, and F1 for different available views on the Flickr photos and 500PX datasets.

	Flickr photos			500PX		
Method	RC	PR	F1	RC	PR	F1
G	20.1	17.3	18.6	16.0	13.7	14.8
V	24.4	19.4	21.6	17.9	16.6	17.2
GT	27.8	22.5	24.9	18.6	16.7	17.6
GV	43.7	43.3	43.5	**35.3**	31.4	33.2
VT	28.1	23.4	25.5	20.5	19.0	19.7
MF	35.9	34.6	35.2	24.7	20.3	22.3
GTV	**45.4**	**44.0**	**44.7**	35.2	**32.1**	**33.6**

The performance of the proposed clustering algorithm in the annotation task is compared with several graph-based multi-view clustering methods in Table 4. In other words, we substitute different graph-based clustering models for the proposed clustering algorithm. The following conclusions can be drawn based on these results.

- Comparing the standard SC performance on different views, it is clear that they achieve different results, which confirm the heterogeneity of available views. Thus, it is vital to distinguish views in constructing a multi-view learning model, as we do in the proposed method with the parameters η-s and q-s.
- All multi-view graph learning methods outperform the single-view graph learning results significantly. This confirms the fact that integrating different views can improve clustering performance.
- Comparing G and V in Table 3 with SC(1) and SC(2) in Table 4, the proposed MVRSC method achieves better performance. This is mainly because of learning a more robust and accurate graph in the proposed method. Remember that we utilize the MCC formulation to construct the similarity graph.
- Moreover, the proposed multi-view graph-based clustering model outperforms other multi-view clustering methods in all datasets. It validates the effectiveness of the proposed graph construction in spectral clustering, proving that combining different views via a robust measure can enhance the clustering performance.

Table 5 shows the comparison of the proposed method and some state-of-the-art multi-view annotation models on the Corel5K dataset. We can see that MVRSC achieves performance improvement of 2.3% in F1. Also, the proposed method surpasses [36,42,39,30,39] in terms of Precision and Recall respectively.

Fig. 5 shows the impact of variable "q" of the proposed method in detecting outlier images in three exemplary concepts (i.e., "street", "fashion", and "ship") in visual view. The proposed method tries to omit the outlier images of each concept by assigning the small value for their "q_{2i}", index 2 indicates visual view (see Table 2). For instance, in the concept of the "ship", since there is no outlier, and the proposed model succeeds in omitting outliers, images with the least "q" in this cluster are relevant to the "ship" concept. However, our method could

Table 4
Clustering performance comparison on all datasets (%).

	Flickr photos			500PX			Corel5K		
Method	RC	PR	F1	RC	PR	F1	RC	PR	F1
SC(1)	18.7	17.6	18.1	16.1	13.3	14.6	27.2	23.6	25.3
SC(2)	22.7	18.9	20.6	17.9	16.0	16.9	19.8	17.1	18.4
SC(3)	–	–	–	–	–	–	25.9	24.7	25.3
MMSC [44]	37.3	28.7	32.4	21.1	17.9	19.4	26.8	26.6	26.7
AMGL [45]	36.9	26.1	30.6	20.7	17.0	18.7	30.0	25.1	27.3
MLAN [43]	**45.4**	26.4	33.4	23.6	19.1	21.1	38.4	38.3	38.3
AWP [46]	40.6	32.8	36.3	27.7	23.2	25.3	43.8	37.6	40.5
GMC[48]	43.5	39.3	41.3	31.8	29.7	30.7	49.7	40.0	44.3
GFSC[49]	43.7	40.2	41.9	30.1	27.9	29.0	51.1	40.5	45.2
Method in [50]	45.1	35.9	40.0	26.3	23.8	25.0	46.4	38.5	42.1
COMIC[51]	39.2	32.7	35.7	25.9	23.0	24.4	43.5	36.7	39.8
MVRSC	**45.4**	**44.0**	**44.7**	**35.2**	**32.1**	**33.6**	**54.3**	**42.9**	**47.9**

Table 5
Comparison of MVRSC and other methods in terms of PR, RC and $F1$ on the Corel5K dataset

Method	RC	PR	F1
NMF-KNN [36]	**56.0**	38	45.2
OPSL [42]	55.0	38.3	45.2
MVSAE [39]	47	37	42
MvNMF [30]	44	**47.5**	45.6
MVRSC	54.3	42.9	**47.9**

not entirely be successful in other concepts, and there are some outliers.

Fig. 7 illustrates the locations of three above-mentioned concepts on three maps[3] with their longitude. Locations of each concept are divided into two categories: dark-blue and red pins, which show the locations of data and outliers (i.e., data with the least "q_{1i}", index 1 indicates geographical view (see Table 2), detected by the proposed method according to their geotags) respectively. For each concept, six locations with the least "q_{1i}" have been shown. It is obvious that most of the outliers are detected and are assigned with the least value of "q" correctly.

To compare the impact of different distances in the tag prediction phase, we evaluate each distance and depict the results in Fig. 6 for the Flickr Photos and 500PX datasets. We can see from the line graphs that fusion distance obtains competitive results compared to the other distances, and these graphs verify its superiority. Moreover, geo-visual distance achieves better performance in comparison with visual and geographical distances.

Fig. 8 depicts some random image annotation examples obtained by the proposed MVRSC on the Flickr photos and 500PX datasets. MVRSC predicts relevant tags in most cases, which show the effectiveness of the proposed method. In some examples (e.g., row 1, col 4, and row 3, col 4), the proposed method retrieves tags that are completely matched with the ground truth. On the other hand, we observe some tags which are not matched with the ground truth but describe the content of images. Specifically, "*nature*" and "*water*" are the relevant tags with the content of the image (row 1, col 5). Also, tag "*canada*" in the image (row 2, col 3) is another one that describes the location of its image but is placed in mismatching tags. This problem is due to the lack of perfect ground truth tags in the real datasets. In total, our method achieves a significant result in the annotation task.

[3] Maps are visualized by https://www.espatial.com/.

5. Conclusion

In this paper, we proposed MVRSC, a multi-view robust spectral clustering method, which tries to realize the semantic concepts of images based on feature-level and decision-level fusion of different views. In the training phase, we derived an optimization problem based on the MCC to construct the relationships between training images and their tags. It tries to omit the outlier data by providing the least "q" for each view in an iterative algorithm. In the test phase, we employ a decision-level fusion method to predict appropriate tags for unknown test data. The stability condition and bound calculation of the proposed method are investigated, as well. Experiments are conducted on the two real-world datasets and corel5k. The experimental results verify the effectiveness of the proposed method in comparison with other state-of-the-art multi-view clustering methods.

For future works, we plan to improve this work in the following directions. First, we intend to consider diversity and redundancy among various views to promote complementary information and improve the performance of the image annotation model. Second, the proposed method may be combined with other useful metadata for the Flickr photos and 500PX datasets, such as title or user information, for better performance. Third, different optimization problems can be investigated to detect the outlier data. Besides, we may adjust weighting for different views in the tag prediction phase in order to fuse them based on the voting method.

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 Proposition 1 comprehensively. HQ modeling is described based on the conjugate function theory [66,67]. The conjugate function $\phi^*(s)$ of a convex function $\phi(p) = -p\log(-p) + p$, where $p<0$, is defined as follows:

$$\begin{aligned}\phi^*(s) &= \sup_{p\in dom\phi}\left(s^T p - \phi(p)\right)\\ &= \sup_{p\in dom\phi}\left(s^T p + p\log(-p) - p\right).\end{aligned} \tag{A.1}$$

Fig. 5. Example of three concepts ("*street*", "*fashion*", and "*ship*") from the Flickr photos dataset with their landmarks based on a visual view. Images in green and red rectangle depict the data that are assigned the least "q_{2i}"- index 2 indicates visual view (see Table 2), - according to their visual view correctly and incorrectly respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

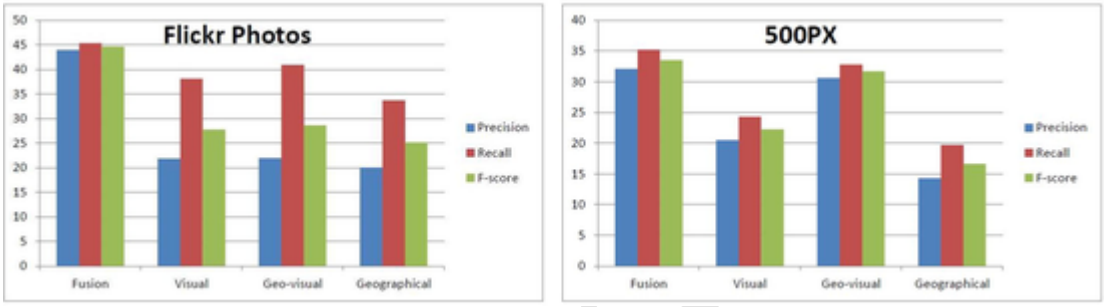


Fig. 6. Precision, Recall, and F-score of MVRSC for different distances in predicting tags on the Flickr Photos and 500PX datasets.

Since the above function is concave with respect to p, we can solve it by setting the derivative of $s^T p + p\log(-p) - p$ to 0 as follows:

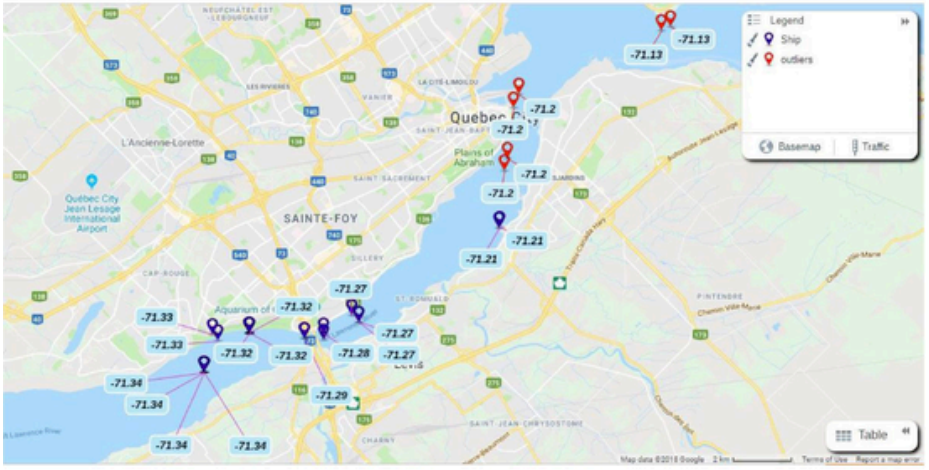
$$\begin{aligned} &\frac{\partial\{s^T p + p\log(-p) - p\}}{\partial p} = 0 \\ &\rightarrow s + \log(-p) = 0 \\ &\rightarrow p = -\exp(-s) < 0. \end{aligned} \tag{A.2}$$

Thus, the supremum is achieved at $p = -\exp(-s)$. Consequently, by replacing p in Eq. (A.1), the conjugate function of $\phi(p)$ is:

$$\phi^*(s) = \exp(-s) \tag{A.3}$$

(a) Map for the "*street*" concept

(b) Map for the "*fashion*" concept



(c) Map for the "*ship*" concept

Fig. 7. Example of three different concepts ("*street*", "*fashion*", and "*ship*") from the Flickr photos dataset according to their geographical view. Coordinates in red and dark-blue pins depict the outliers (coordinates with the least "g_{i1}", where index 1 indicates geographical view (see Table 2)) and other coordinates respectively.

Fig. 8. Annotation examples on the Flickr photos and 500PX datasets. The first two rows and the last ones indicate the examples from Flickr photos and 500PX datasets respectively. Ground truth is the tags in black given by the dataset. The tags in color are predicted by the proposed MVRSC, where the matched tags are shown in green, and the red ones are mismatching tags in the ground truth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

[1] K. Jiang, H. Yin, P. Wang, N. Yu, Learning from contextual information of geo-tagged web photos to rank personalized tourism attractions, Neurocomputing 119 (2013) 17–25.

[2] H. Ma, S.A. Ay, R. Zimmermann, S.H. Kim, Large-scale geo-tagged video indexing and queries, Geoinformatica (2014) 671–697.

[3] Y. Lu, C. Shahabi, S.H. Kim, Efficient indexing and retrieval of large-scale geo-tagged video databases, Geoinformatica (2016) 829–857.

[4] X. Qian, X. Liu, C. Zheng, Y. Du, X. Hou, Tagging photos using users' vocabularies, Neurocomputing 111 (2013) 144–153.

[5] R. Abbasi, M. Grzegorzek, S. Staab, Large scale tag recommendation using different image representations, in: Semantic Multimedia, 4th International Conference on Semantic and Digital Media Technologies, 2009, pp. 65–76.

[6] I. Miliou, A. Vlachou, Location-aware tag recommendations for flickr, International Conference on Database and Expert Systems Applications, 2014, pp. 97–104.

[7] S.S. Lee, D. Won, D. McLeod, Tag-geotag correlation in social networks, in: Proceedings of ACM workshop on Search in social media, 2008.

[8] K. Toyama, R. Logan, A. Roseway, P. Anandan, Geographic location tags on digital images, ACM International Conference on Multimedia, 2003, pp. 156–166.

[9] W. Liu, J. Wang, A.K. Sangaiah, J. Yin, Dynamic metric embedding model for point-of-interest prediction, Fut. Gen. Comput. Syst. 83 (2018) 183–192.

[10] X. Ren, M. Song, H. E, J. Song, Context-aware probabilistic matrix factorization modeling for point-of-interest recommendation, Neurocomputing 241 (2017) 38–55.

[11] G. Cai, K. Lee, I. Lee, Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos, Expert Syst. Appl. 94 (2017) 32–40.

[12] E. Spyrou, P. Mylonas, Analyzing flickr metadata to extract location-based information and semantically organize its photo content, Neurocomputing 172 (2016) 114–133.

[13] S. Ashkezari-Toussi, M. Kamel, H. Sadoghi-Yazdi, Emotional maps based on social networks data to analyze cities emotional structure and measure their emotional similarity, Cities 86 (2019) 113–124.

[14] J. Hays, A.A. Efros, Im2gps: estimating geographic information from a single image, in: Proceeding of ieee conference on computer vision and pattern recognition (CVPR), 2008.

[15] X. Li, M. Larson, A. Hanjalic, Global-scale location prediction for social images using geo-visual ranking, IEEE Trans. Multimedia 17 (2015) 674–686.

[16] L. Zheng, Z. Caiming, C. Caixian, MMDF-LDA: An improved multi-modal latent dirichlet allocation model for social image annotation, Expert Syst. Appl. 104 (2018) 168–184.

[17] Q. Cheng, Q. Zhang, P. Fu, C. Tu, S. Li, A survey and analysis on automatic image annotation, Pattern Recogn. (2018) 242–259.

[18] M. Sangeetha, K. Anandakumar, A. Bharathi, Automatic image annotation and retrieval: a survey, Int. Res. J. Eng. Technol. (IRJET) (2016).

[19] F.M. Belem, J.M. Almeida, M.A. Gonçalves, A survey on tag recommendation methods, J. Assoc. Inf. Sci. Technol. 68 (2016) 830–844.

[20] J. Luo, D. Joshi, J. Yu, A. Gallagher, Geotagging in multimedia and computer vision—a survey, Multimed Tools Appl. 51 (2011) 187–211.

[21] Y.-T. Zheng, Z.-J. Zha, T.-S. Chua, Research and applications on georeferenced multimedia: a survey, Multimedia Tools Appl. 51 (2011) 77–98.
[22] C. Lei, D. Liu, W. Li, Social diffusion analysis with common-interest model for image annotation, IEEE Trans. Multimedia 18 (2016) 687–701.
[23] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. (2007) 395–416.
[24] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, E.Y. Chang, Parallel spectral clustering in distributed systems, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 568–586.
[25] L. He, N. Ray, Y. Guan, H. Zhang, Fast large-scale spectral clustering via explicit feature mapping, IEEE Trans. Cybernet. 49 (2019) 1058–1071.
[26] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using crossmedia relevance models, in, Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 119–126.
[27] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple bernoulli relevance models for image and video annotation, Computer Vision and Pattern Recognition (CVPR), 2004.
[28] D. Putthividhy, H. Attias, S. Nagarajan, Topic regression multi-modal latent dirichlet allocation for image annotation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010).
[29] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, H.-J. Zhang, A probabilistic semantic model for image annotation and multimodal image retrieval, IEEE International Conference on Computer Vision (ICCV), 2005.
[30] R. Rad, M. Jamzad, Image annotation using multi-view non-negative matrix factorization with different number of basis vectors, J. Vis. Commun. Image Represent. 46 (2017) 1–12.
[31] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, G. Mori, Learning structured inference neural networks with label relations, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2960–2968.
[32] P. Savita, D. Patel, A. Sinhal, A neural network approach to improve the efficiency of image annotation, Int. J. Eng. Res. Technol. (IJERT) (2013) 35–41.
[33] Y. Verma, C.V. Jawahar, Exploring svm for image annotation in presence of confusing labels, Proceedings British Machine Vision Conference, 2013.
[34] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, Proceedings of the 10th European conference on computer vision, 2008, pp. 316–329.
[35] Y. Verma, C.V. Jawahar, Image annotation using metric learning in semantic neighborhoods, European Conference on Computer Vision (ECCV), 2012.
[36] M.m. Kalayeh, H. Idrees, M. Shah, NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[37] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation, IEEE 12th International Conference Computer Vision, 2009.
[38] R. Rad, M. Jamzad, Automatic image annotation by a loosely joint non-negative matrix factorisation, IET Comput. Vision (2015) 806–813.
[39] Y. Yang, W. Zhang, Y. Xie, Image automatic annotation via multi-view deep representation, J. Vis. Commun. Image Represent. (2015) 368–377.
[40] V.N. Murthy, E.F. Can, R. Manmatha, A hybrid model for automatic image annotation, Proceedings of International Conference on Multimedia Retrieval, 2014.
[41] M. Zhao, T.W. Chow, Z. Zhang, B. Li, Automatic image annotation via compact graph based semi-supervised learning, Knowl.-Based Syst. 76 (2015) 148–165.
[42] Z. Xue, G. Li, Q. Huang, Joint multi-view representation and image annotation via optimal predictive subspace learning, Inf. Sci. (2018) 190–194.
[43] F. Nie, G. Cai, J. Li, X. Li, Auto-weighted multi-view learning for image clustering and semi-supervised classification, IEEE Trans. Image Process. (2017) 1501–1511.
[44] X. Cai, F. Nie, H. Huang, F. Kamangar, Heterogeneous image feature integration via multi-modal spectral clustering, Computer Vision and Pattern Recognition (CVPR), 2011.
[45] F. Nie, J. Li, X. Li, Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification, International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1881–1887.
[46] F. Nie, L. Tian, X. Li, Multiview clustering via adaptively weighted procrustes, KDD (2018).
[47] Z. Hu, F. Nie, W. Chang, S. Hao, R. Wang, X. Li, Multi-view spectral clustering via sparse graph learning, Neurocomputing 384 (2020) 1–10.
[48] H. Wang, Y. Yang, B. Liu, GMC: Graph-based multi-view clustering, IEEE Trans. Knowl. Data Eng. 32 (2019) 1116–1129.
[49] Z. Kang, G. Shi, S. Huang, W. Chen, X. Pu, J.T. Zhou, Z. Xu, Multi-graph fusion for multi-view spectral clustering, Knowl.-Based Syst. 189 (2020) 105102.
[50] C. Tang, X. Zhu, X. Liu, M. Li, P. Wang, C. Zhang, L. Wang, Learning a joint affinity graph for multiview subspace clustering, IEEE Trans. Multimedia 21 (2019) 1724–1736.
[51] X. Peng, Z. Huang, J. Lv, H. Zhu, J.T. Zhou, COMIC: Multi-view clustering without parameter selection, Int. Conf. Mach. Learn. 97 (2019) 5092–5101.
[52] J. Wen, Z. Zhang, Z. Zhang, L. Fei, M. Wang, Generalized incomplete multiview clustering with flexible locality structure diffusion, IEEE Trans. Cybernet. (2020) 1–14.
[53] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, E. Zhu, Efficient and effective regularized incomplete multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1.
[54] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, W. Gao, Late fusion incomplete multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 2410–2423.
[55] W. Liu, P.P. Pokharel, J.C. Principe, Correntropy: properties and applications in non-gaussian signal processing, IEEE Trans. Signal Process. 55 (2007) 5286–5298.
[56] M. Maier, U. von Luxburg, M. Hein, How the result of graph clustering methods depends on the construction of the graph, ESAIM: Probability and Statistics, pp. 370–418.
[57] N. Zhou, H. Cheng, J. Qin, Y. Du, B. Chen, Robust high-order manifold constrained sparse principal component analysis for image representation, IEEE Trans. Circuits Syst. Video Technol. 29 (2018) 1946–1961.
[58] R. He, B.-G. Hu, W.-S. Zheng, X.-W. Kong, Robust principal component analysis based on maximum correntropy criterion, IEEE Trans. Image Process. 20 (2011) 1485–1494.
[59] N. Zhou, Y. Xu, H. Cheng, Z. Yuan, B. Chen, Maximum correntropy criterion based sparse subspace learning for unsupervised feature selection, IEEE Trans. Circuits Syst. Video Technol. 29 (2017) 404–417.
[60] R. He, W.-S. Zheng, B.-G. Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1561–1576.
[61] B. Chen, L. Xing, H. Zhao, N. Zheng, J.C. Principe, Generalized correntropy for robust adaptive filtering, IEEE Trans. Signal Process. 64 (2016) 3376–3387.
[62] B. Chen, L. Xing, J. Liang, N. Zheng, J.C. Principe, Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion, IEEE Signal Process. Lett. 21 (2014) 880–884.
[63] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 95–125.
[64] G. Liu, Z. Lin, S. Yan, J. Sun, Y.M.Y. Yu, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. (2013) 171–184.
[65] R. Vidal, Subspace clustering, IEEE Signal Process. Mag. 28 (2011) 52–68.
[66] R. He, W.S. Zheng, T. Tan, Z. Sun, Half-quadratic-based iterative minimization for robust sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 261–275.
[67] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, New York, USA, 2004.
[68] Sumuya, C. Guo, S. Chai, A note on spectral clustering method based on normalized cut criterion, Chinese Conference on Pattern Recognition (CCPR), 2009.
[69] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[70] E.E. far, R. Vidal, Sparse manifold clustering and embedding, Neural Inf. Process. Syst. (NIPS) (2011).
[71] P. Duygulu, K. Barnard, J. de Freitas, D. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, Computer Vision, ECCV, Springer, 2002, pp. 97–112.
[72] C.D. Manning, P. Raghavan, H. Schutze, Introduction to information retrieval, Cambridge University Press, 2008.
[73] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Neural Inf. Process. Syst. (NIPS) (2001) 849–856.
[74] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (NIPS) (2012).
[75] J. Deng, W. Dong, R. Socher, Imagenet: A large-scale hierarchical image database, IEEE Comput. Vision Pattern Recogn. (CVPR) (2009).
[76] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Int. J. Inf. Process. Manage. 24 (1988) 513–523.
[77] D.G. Lowe, Object recognition from local scale-invariant features, Comput. vision (1999) 1150–1157.