Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# Shared-private Information Bottleneck Method for Cross-modal Clustering

**XIAOQIANG YAN[1], YANGDONG YE[1] (MEMBER, IEEE), YIQIAO MAO[1], AND HUI YU[2], (Senior member, IEEE)**

[1]School of Information Engineering, Zhengzhou University, Zhengzhou, 450052, China (e-mail: iexqyan@zzu.edu.cn)
[2]School of Creative Technologies, University of Portsmouth, PO1 2DJ, United Kingdom (e-mail: hui.yu@port.ac.uk)

Corresponding author: Yangdong Ye (e-mail: ieydye@zzu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61772475, in part by the National Key Research and Development Program of China under Grant 2018YFB1201403.

**ABSTRACT** Cross-modal analysis has recently drawn much attention due to the rapid growth and widespread emergence of multimodal data. It integrates multiple modalities to improve the learning and generalization performance. However, most previous methods just focus on learning a common shared feature space for all modalities and ignore the private information hidden in each individual modality. To address this problem, we propose a novel shared-private information bottleneck (SPIB) method for cross-modal clustering. Firstly, we devise a hybrid words model and a consensus clustering model to construct the shared information of multiple modalities, which capture the statistical correlation of low-level features and the semantic relations of the high-level clustering partitions, respectively. Secondly, the shared information of multiple modalities and the private information of individual modalities are maximally preserved through a unified information maximization function. Finally, the optimization of SPIB function is performed by a sequential "draw-and-merge" procedure, which guarantees the function converge to a local maximum. Besides, to solve the lack of tags in cross-modal social images, we also investigate the use of structured prior knowledge in the form of knowledge graph to enrich the information in semantic modality, and design a novel semantic similarity measurement for social images. Experimental results on four types of cross-modal datasets demonstrate that our method outperforms the state-of-the-art approaches.

**INDEX TERMS** Cross-modal clustering, information bottleneck, mutual information, knowledge graph, social images.

## I. INTRODUCTION

WITH the rapid advancement of information technologies, massive amounts of cross-modal multimedia data are rapidly generated on the Internet over the last decade. It is common that the cross-modal data have similar high-level semantic information but appear in different modalities, sources, spaces. As shown in Fig. 1, a piece of news can be reported in different languages [1]; One image can be characterized by heterogeneous shape, texture and color features [2]; A concept or event can be revealed by multiple media types, such as image, text, video and audio. There also have been various applications for cross-modal data, such as cross-modal retrieval [3]–[5], multi-source fusion [6], [7], cross-modal hashing [8], multi-sensors surveillance [9]. However, most of these applications rely on the availability of a large number of labeled data to train a learning model. Thus, it is natural to resort $clustering$ algorithms for mining the undiscovered knowledge in unlabeled cross-modal data.

Clustering is defined as an unsupervised learning method [10], [11] where the similar objects are partitioned into together while separating dissimilar ones apart. Although the existing clustering approaches have demonstrated superior performance on various tasks, they cannot directly deal with cross-modal data due to its multi-source and heterogeneous characteristic, which has been widely considered as a great challenge to cross-modal analysis. Therefore, it is urgent to develop a cross-modal clustering (CMC) algorithm so as to integrate multiple content modalities.

Basically, CMC [12], [13] is a type of data-driven analysis method, which aims to find a more reasonable cluster structure and reveal the hidden relationships between multiple modalities by considering the complementary effect of the different modalities. Obviously, the key issue of CMC methods is to capture the relatedness of different modalities and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



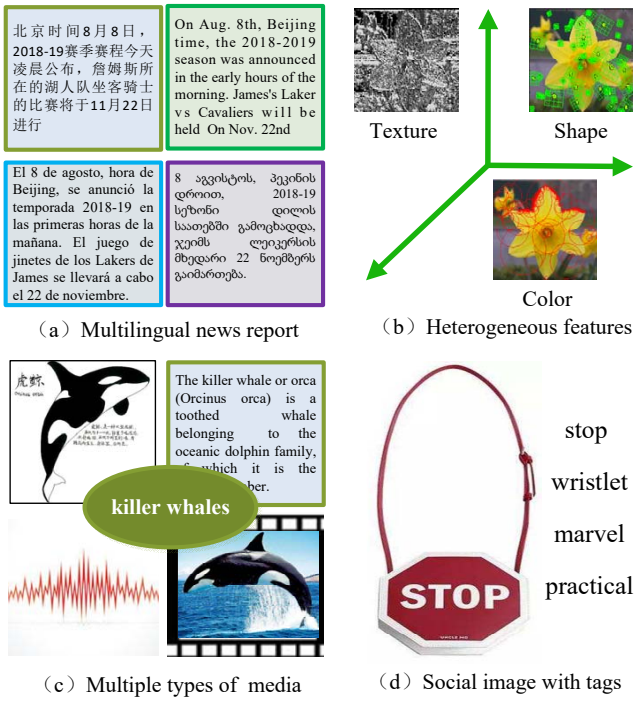


FIGURE 1: The cases of cross-modal data. a) A piece of news in different languages. b) Heterogeneous descriptors of one image. c) A character is profiled by multiple media types. d) one social image with several tags.

transfer the shared information across modalities. To this end, the most prevalent methods are to learn a latent shared space by maximizing the correlations between multiple modalities, such as canonical correlation analysis (CCA) [12], [14], shared kernel information embedding (KIE) [13], Gaussian process latent variable model [15]. However, the subspace based CMC approaches will destroy the original feature structure of the cross-modal data and lead to the loss of some necessary information when the features of all modalities are mapped into a shared low-dimensional space. In addition to the subspace approaches, there are some other effective clustering strategies that can be applied to cluster cross-modal data, such as multi-modal latent Dirichlet allocation [16], multi-modal spectral clustering [2], multi-view nonnegative matrix factorization [17], robust multi-view $k$-means [18]. However, all the existing CMC approaches just focus on the shared information of the multiple modalities and ignore the private information hidden in each individual modality, which is obviously unreasonable and not in conformity with the realistic applications.

Recently, the popularity of social media websites has successfully motivated users to tag their visual content on the web, which results in large massive of cross-modal social images [19]–[21]. The social images include two modalities: image and user tags. Usually, the most users upload images with very simple words or tags, which leads to rare textual information in social images. Moreover, the tags are often ambiguous, overly personalized and limited in terms of completeness [22]. This is not surprising because that the diversity of knowledge and cultural background of its users results the uncontrolled nature of social tagging. Besides, there is also no clear semantic relationships between the social tags, which makes the clustering of cross-modal social images still a challenging task.

To address these problems, we propose a novel shared and private information bottleneck (SPIB) method for cross-modal clustering (see Fig 2). Firstly, to construct the shared information of multiple modalities, we devise a hybrid words model and a consensus clustering model, which can characterize the shared information from the correlations of low-level features and the semantic relations of the high-level clustering partitions, respectively. Then, a unified information maximization objective function is proposed to maximally preserve the shared information of multiple modalities and the private information of individual modalities. Finally, a sequential "draw-and-merge" procedure is proposed to optimize the SPIB objective function. Besides, to solve the lack of semantic information in cross-modal social images, we also investigate the use of structured prior knowledge in the form of knowledge graph, and design a novel semantic similarity measurement for social images. Experimental results on cross-modal clustering tasks demonstrate that our method outperforms the state-of-the-art approaches. The contributions of this study can be summarized as follows.

- A new cross-modal clustering method named SPIB is proposed, which comprehensively considers the shared information of multiple modalities and the private information of each individual modality.
- Two shared information construction methods, hybrid words model and consensus clustering model, are proposed, which can ensure the statistical correlation of low-level features and the semantic relations of the high-level clustering partitions, respectively.
- To solve the lack of semantic information in cross-modal social images, we investigate the use of structured prior knowledge in the form of knowledge graph, and design a novel semantic similarity measurement for social cross-modal data.

The remainder of this paper is organized as follows. Section II briefly reviews the related work on cross-modal analysis, then gives the basic background of cross-modal clustering and information bottleneck. Then the proposed shared-private information bottleneck method is presented in Section 3. The experimental results are detailed in Section 4. Finally, Section 5 concludes the paper.

## II. RELATED WORK

### A. CROSS-MODAL CLUSTERING

Cross-modal clustering [2], [12]–[15], [17], [18] approaches aim to find more reasonable cluster structure and reveal the hidden relationships between multiple modalities by considering the complementary effect of the different modalities.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



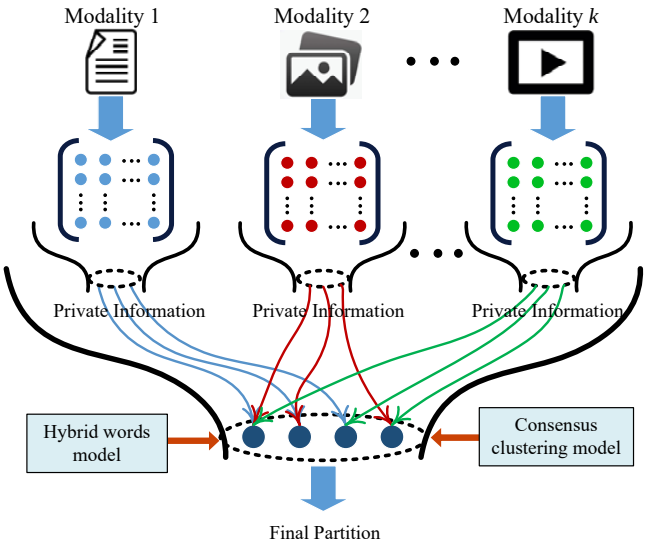


FIGURE 2: The illustration of SPIB method. In the SPIB model, the shared information in each modality and the shared information constructed by hybrid words model and consensus model are maximally preserved through "bottleneck", which is performed by a unified information maximization objective function based on information bottleneck method.

To this end, the most prevalent methods are to learn a latent shared space which maximizes the correlations between multiple modalities. For instance, [13]–[15] utilize shared kernel information embedding, canonical correlation analysis, Gaussian process latent variable models respectively to construct the shared subspace of multiple modalities. However, the subspace based CMC approaches always destroy the original feature structure of the cross-modal data and lead to the loss of some necessary information when the features of all modalities are mapped into shared a low-dimensional space. Aiming at this problem, [23] proposes a hierarchical model to combine the text and visual features in a bottom-up manner. [24] first segments the target image into different patches, then the image and text data is transformed into a cross-modal vector by latent Dirichlet allocation model. From the methodological point of view, [2], [17], [18] propose multi-model spectral clustering, multi-view joint matrix factorization and robust multi-view k-means method to deal with the cross-modal data. However, all the aforementioned approaches just focus on the shared information of the multiple modalities and ignore the private information hidden in each individual modal, which is obviously unreasonable and not in conformity with the realistic applications.

Consensus clustering [25]–[28], also named ensemble clustering, has drawn more and more attention since it naturally has the ability to leverage complementary information from heterogeneous data sources. In particular, when dealing with cross-modal data, the consensus clustering method first generates a set of basic partitions for each individual modality, then the multiple basic partitions are integrated into a consensus one by a given combination criterion, such as shared mutual information [25], probability trajectories [29], cluster uncertainty estimation [30]. However, the existing consensus clustering methods just rely on the quality of basic partitions and ignores the original feature distribution of the original data, which always leads to the over-reliance of final clustering partition on the quality of the basic partitions. Recently, [26] proposes a consensus information bottleneck (CIB) method, which can simultaneously deal with the original features and the basis clusterings to relief the over-reliance of the existing consensus clustering method on the quality of basic clusterings. However, CIB method copes with each original features independently and ignores the low-level statistical correlations. Besides, the CIB method focuses on the multi-feature scenario of single-modal data, and cannot effectively cope with the problem of cross-modal clustering. In contrast, we devise a hybrid words model and a consensus clustering model in the proposed SPIB method to construct the shared information of multiple modalities more comprehensively, which simultaneously ensure the statistical correlation of low-level features and the semantic relations of the high-level clustering partitions.

### B. INFORMATION BOTTLENECK

Information bottleneck (IB) [31] is a typical data analysis method based on information theory. Suppose there is the joint distribution between dataset $X$ and its feature variable $Y$, IB aims to find an optimal and compressed representation $T$ for the source variable $X$, meanwhile, the information contained in feature variable $Y$ is maximally preserved. The objective function of IB method is given by

$$\mathcal{R}(D) = \min_{\{p(t|x):I(T;Y)\leq D\}} I(T;X), \tag{1}$$

where $p(t|x)$ is the coding scheme between the source variable $X$ and its compressed variable $T$, $I(T;X)$ is the mutual information measuring the shared information between two variables. As shown in Eq (1), IB method aims to find an optimal coding scheme $p(t|x)$ under the condition $I(T;Y) \leq D$, where $D$ is the set of all the possible coding schemes between $X$ and $T$. Thus, the objective function of IB is rewritten as

$$\mathcal{L}_{max}[p(t|x)] = I(T;Y) - \beta^{-1} I(T;X), \tag{2}$$

where $\beta$ balances the trade-off between the data compression $I(T;X)$ and information preservation $I(T;Y)$. In clustering scenario, the size of source dataset $X$ is always much larger than its compressed representation $T$, i.e., the cluster structure. Thus, $\beta$ is always set as $\infty$ in real world applications, such as document analysis [32], object recognition [33], [34], action clustering [35], [36]. Thus, the objective function of IB is simplified as

$$\mathcal{L}_{max}[p(t|x)] = I(T;Y). \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



## III. SHARED-PRIVATE INFORMATION BOTTLENECK FOR CROSS-MODAL CLUSTERING

In this section, we elaborate a novel shared-private information bottleneck method, which establishes a general framework of cross-modal clustering. First, we give the problem formulation of the proposed SPIB method. Second, two shared information construction models are presented. third, the objective function of the SPIB and its optimization method are provided. Then, we give the theoretical analysis of the SPIB method. Finally, the semantic extension model is proposed to solve the lack of tags in social images.

### A. PROBLEM FORMULATION

Suppose a cross-modal dataset is composed of $k$ modalities for $n$ instances, which is denoted by a set of matrices $(X, Y^1), \cdots, (X, Y^k)$, where the variable $X = \{x_1, \cdots, x_n\}$ denotes the set of data instances, the variables $Y^1, \cdots, Y^k$ denote the feature vectors of the $k$ modalities. In this study, the feature variables $Y^1, \cdots, Y^k$ in each modality are treated as private information, while the variable $S$ is defined to indicate the shared information between the $k$ modalities. Thus, the overall objective function of SPIB method is given as the following definition.

**Definition 1.** *(Overall objective function of the SPIB). Suppose there is a cross-modal dataset $(X, Y^1), \cdots, (X, Y^k)$, if the private information $Y^1, \cdots, Y^k$ and shared information $S$ are given, the overall objective function of the SPIB is defined as*

$$\mathcal{L}_{max}[p(t|x)] = \sum_{i=1}^{k} I(T; Y^i) + \lambda \cdot I(T; S), \qquad (4)$$

*where $p(t|x)$ is the mapping from source dataset $X$ to its cluster structure $T$. $\sum_{i=1}^{k} I(T; Y^i)$ are the mutual information [37] between $T$ and private information $Y^1, \cdots, Y^k$. $I(T; S)$ is the mutual information between $T$ and shared information $S$. $\lambda$ is the trade-off parameter to balance the shared and private information.*

As shown in the Eq. (4), the SPIB method aims to discover the optimal cluster structure $T$ of the cross-modal dataset $X$, while the shared and private information are maximally preserved with respect to $T$. When $\lambda = 0$ and $k = 1$, the SPIB reduces to IB algorithm. In the SPIB framework, we can employ bag-of-words model [38] or bag-of-visual-words model [39] to construct the joint distributions $p(X, Y^1), \cdots, p(X, Y^k)$ . Thus, the private information term $\sum_{i=1}^{k} I(T; Y^i)$ is computable. Next, we elaborate two models to construct the shared information of multiple modalities.

### B. SHARED INFORMATION CONSTRUCTION

#### 1) Hybrid Words Model

Bag-of-words or bag-of-visual-words model [38], [39] is a popular data representation technique, which represents the source data as the co-occurrent vectors of keywords or visual words. For instance, a piece of news about basketball can be described by the occurrence number of the words like score, coach, fast break, rebound, etc; A urban image can be characterized by the occurrence number of the words like building, street, traffic light, etc. Simply combining the co-occurrent vectors of multiple modalities can depict the relationship between different modalities, however, the word frequency vectors of different modalities have obvious differences in scale and large sample redundancy. Aiming at this problem, we propose a hybrid words model to capture the relationship of multiple modalities. First, the data instances in each modality are transformed into co-occurrence vectors independently. Then, a novel hierarchical information maximization (HIM) method is proposed to build the hybrid words space of all modalities (see Fig. 3). Thus, the low-level statistic similarity of different modalities can be maximally ensured.

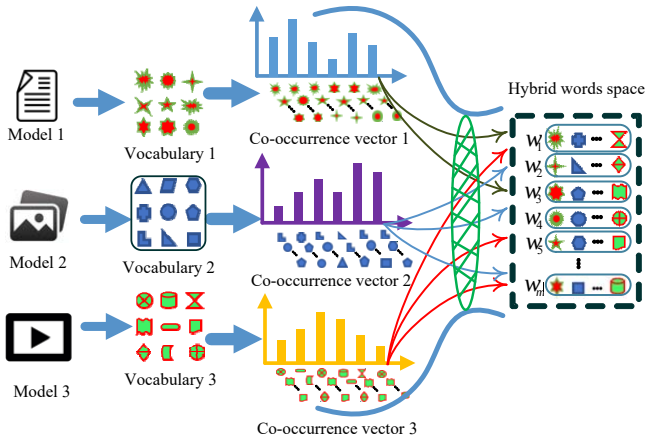


FIGURE 3: The flowchart of hybrid words model. In the hybrid words model, the data instances in each modality are firstly transformed into co-occurrent vectors independently. Then, a novel hierarchical information maximization method is proposed to build the hybrid words space of all modalities.

Suppose a cross-modal dataset consists of $k$ modalities, and we use $k$ feature variables $Y^1, \cdots, Y^k$ to indicate the $k$ modalities, where $Y^i$ is the feature variable of the $i$-th modality. The $Y^i$ takes values from $Y^i = \{y_1^i, \cdots, y_{m_i}^i\}$, where $y_j^i$ denotes the occurrence number of word $y_{x_j}^i$ in data $x_j$, and $m_i$ is the vocabulary size in the $i$-th modality. Thus, we define the following function to discover the hybrid words space $\tilde{Y}$ for all modalities.

$$\mathcal{F}_{max}[p(\tilde{y}|y^i)] = I(\bar{Y}; X) - \sum_{i=1}^{k} I(\tilde{Y}; Y^i), \qquad (5)$$

where $p(\bar{y}|y^i)$ is the mapping from the feature vector $Y^i$ of each modality to the hybrid words space $\tilde{Y}$.

In this study, we propose a HIM method to optimize the Eq. (5). First, the HIM method regards each feature vector in all modalities as a singleton cluster, i.e., $|\tilde{Y}| = \sum_{i=1}^{k} |Y^i|$, where $|\tilde{Y}|$ is the size of the hybrid words space, $|Y^i|$ is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



the number of features in the $i$-th modality. Then, the HIM method integrates the most similar features together in a bottom-up manner, in which the redundant features are gradually eliminated. Suppose $y_h$ and $y_g$ are two feature vectors, then the value change of Eq. (5) is calculated as

$$\Delta\mathcal{F}_{max}(y_h, y_g) = \Delta\mathcal{F}_{max}^{bef} - \Delta\mathcal{F}_{max}^{aft}, \tag{6}$$

where $\Delta\mathcal{F}_{max}^{bef}$ and $\Delta\mathcal{F}_{max}^{aft}$ are the values of Eq. (5) before and after $y_h$ and $y_g$ are merged together. Then, the probability distribution is defined as following after the $y_h$ and $y_g$ are merged together.

$$\begin{cases} p(\tilde{y}) = p(y_h) + p(y_g), \\ p(y^i|\tilde{y}) = \dfrac{p(y_h)}{p(\tilde{y})} p(y^i|y_h) + \dfrac{p(y_g)}{p(\tilde{y})} p(y^i|y_g). \end{cases} \tag{7}$$

Next, we present the processing procedure of the HIM method as follows:

1) Initialize each feature point into a singleton cluster;

2) Calculate the value changes $\Delta\mathcal{F}_{max}(y_h, y_g)$ of Eq. (6);

3) Merge all the pairs of features that satisfy the minimum value change of Eq. (6), i.e., $\arg\min \Delta\mathcal{F}_{max}(y_h, y_g)$;

4) Update $p(\tilde{y})$, $p(y^i|\tilde{y})$ according to Eq. (7).

### 2) Consensus Clustering Model

The proposed hybrid words model characterizes the correlations of multiple modalities by the low-level features. To further explore the relationships of multiple modalities, we propose a consensus clustering model to capture the high-level similarity of the basic clusterings constructed by multiple modalities. In the consensus clustering model, a set of basic clusterings is first constructed from all modalities, then the mutual information metric is adopted to measure the similarity between the basic clusterings of the heterogeneous modalities.

In this study, the proposed SPIB method aims to find the cluster structure $T = \{t_1, \cdots, t_M\}$ in a cross-modal dataset, where $M$ is the number of clusters. Suppose there are $k$ basic clusterings $C = \{C^1, \cdots, C^k\}$ built by the $k$ modalities, in which the $l$-th basic clustering is denoted by $C^l = \{c_1^l, \cdots, c_M^l\}$. To calculate the mutual information $I(T; C^l)$, we should first obtain the joint and margin distribution between them. Suppose $n_i$ is the number of data instances that are allocated into cluster $c_i^l$; $n_j$ is the number of data instances that are allocated into cluster $t_j$; $n_{ij}$ is the number of data instances that are simultaneously allocated into cluster $c_i^l$ and $t_j$. Thus, the probability distributions between cluster $T$ and $C^l$ are calculated as follows

$$p(c_i^l, t_j) = \frac{n_{ij}}{n}, p(c_i^l) = \frac{n_i}{n}, p(t_j) = \frac{n_j}{n}. \tag{8}$$

Fig 4 illustrates the correlation of high-level clusterings between modalities by mutual information metric. In this figure, the black box means the data $x$ appears in the corresponding cluster. From Fig 4. (a) to (c), the degree of similarity between cluster $C^l$ and $T$ is gradually weakened, and the value of mutual information correspondingly decreases.

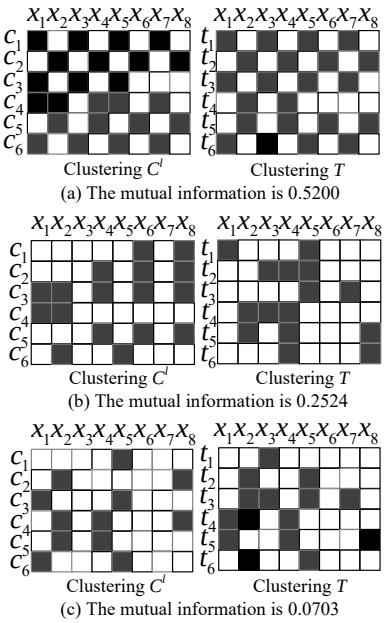


FIGURE 4: The mutual information between the high-level clusterings of different modalities.

This demonstrates that the mutual information can effectively measure the correlation between the high-level clusterings of different modalities.

## C. OBJECTIVE FUNCTION OF SPIB

In this study, the proposed SPIB method aims to discover the optimal cluster structure $T$ of the cross-modal dataset $X$, while the private information $Y^1, \cdots, Y^k$ and shared information $S$ are maximally preserved with respect to $T$. The shared information $S$ includes two parts: the hybrid words space $\tilde{Y}$ and high-level basic clusterings $C = \{C^1, \cdots, C^k\}$, which are constructed by the proposed hybrid words space and consensus clustering model, respectively. Thus, the objective function of the SPIB method is rewritten as

$$\begin{aligned} \mathcal{L}_{max}[p(t|x)] &= \sum_{i=1}^{k} I(T; Y^i) + \lambda \cdot I(T; S) \\ &= \sum_{i=1}^{k} I(T; Y^i) + \lambda \cdot [I(T; \tilde{Y}) + \sum_{i=1}^{k} I(T; C^i)], \end{aligned} \tag{9}$$

where $\sum_{i=1}^{k} I(T; Y^i)$ measures the mutual information between the cluster structure $T$ and private variables $Y^1, \cdots, Y^k$. $I(T; \tilde{Y})$ indicates the amount of information between $T$ and the hybrid words space $\tilde{Y}$. $\sum_{i=1}^{k} I(T; C^i)$ is the correlation of the high-level clusterings of multiple modalities.

## D. OPTIMIZATION OF SPIB

In this part, we propose a sequential "draw-and-merge" to optimize the objective function of the proposed SPIB method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



The sequential "draw-and-merge" approach finds the optimal cluster by the following three steps:

1) **Random partition**. The cross-modal dataset $X = \{x_1, \cdots, x_n\}$ is randomly partitioned into $M$ clusters $T = \{t_1, \cdots, t_M\}$.

2) **Draw**. Every data instance $x$ is drawn from its original cluster $t^{old}$ and treated as a singleton cluster $\{x\}$. Now, the number of clusters is $M$+1.

3) **Merge**. To make the cluster number is $M$, the singleton cluster $\{x\}$ should be allocated into the other cluster $t^{new}$ that will increase the value of Eq. (9).

The core problem of the sequential "draw-and-merge" approach is to select the appropriate cluster $t^{new}$ that the singleton cluster $\{x\}$ is merged into. We use $\mathcal{L}^{bef}$ and $\mathcal{L}^{aft}$ to indicate the value of Eq. (9) before and after $x$ is merged into some new cluster $t^{new}$ that $t^{new} = \arg\min \Delta\mathcal{L} = \mathcal{L}^{bef} - \mathcal{L}^{aft}$, where $\Delta\mathcal{L}$ is the value change of Eq. (9) before and after $x$ is merged into $t^{new}$, and we call it merge cost in this study.

$$\begin{aligned}\Delta\mathcal{L} &= \mathcal{L}^{bef} - \mathcal{L}^{aft} = \sum_{i=1}^{k}[I(T^{bef};Y^i) - I(T^{aft};Y^i)] + \\ &\lambda[I(T^{bef};\tilde{Y}) - I(T^{aft};\tilde{Y}) + \lambda\sum_{i=1}^{k}[I(T^{bef};C^i) - I(T^{aft};C^i)] \\ &= \sum_{i=1}^{k}\Delta I_i^{private} + \lambda\Delta I^{common} + \lambda\sum_{i=1}^{k}\Delta I_i^{clustering}\end{aligned} \tag{10}$$

where $\Delta I_i^{private}$, $\Delta I^{common}$ and $\Delta I_i^{clustering}$ are the merge cost cause by the terms $I(T;Y^i)$, $I(T;\tilde{Y})$ and $I(T;C^i)$, respectively.

**Definition 2.** *Suppose a singleton cluster $\{x\}$ is merged into cluster $t$ and generate a new cluster $\tilde{t}$, we can get*

$$\begin{cases} p(\tilde{t}) = p(x) + p(t), \\ p(y^i|\tilde{t}) = \dfrac{p(x)}{p(\tilde{t})}p(y^i|x) + \dfrac{p(t)}{p(\tilde{t})p(y^i|t)}. \end{cases} \tag{11}$$

Next, we first give the calculation of $\Delta I_i^{private}$ according to Eq.(11) and the definition of mutual information.

$$\begin{aligned}\Delta I_i^{private} &= I(T^{bef};Y^i) - I(T^{aft};Y^i) = \\ &p(t)\sum_{y^i\in Y^i}p(y^i|t)\log\frac{p(y^i|t)}{p(y^i)} + p(x)\sum_{y^i\in Y^i}p(y^i|x)\log\frac{p(y^i|x)}{p(y^i)} \\ &- p(\tilde{t})\sum_{y^i\in Y^i}p(y^i|\tilde{t})\log\frac{p(y^i|\tilde{t})}{p(y^i)}.\end{aligned} \tag{12}$$

**Algorithm 1** The SPIB Algorithm

**Input:**
Different Modalities $p(X,Y^1), \cdots, p(X,Y^k)$.
Parameter $\lambda$.
Cluster number $M$.
**Output:** The cluster assignment $p(t|x)$.
**Preprocessing:**
Generate the hybrid words space $\tilde{Y}$, and the size of $|\tilde{Y}|$ is identified by $|\tilde{Y}| = \frac{|Y^1|+\cdots+|Y^k|}{k}$.
Generate the basic clusterings $C = \{C^1, \cdots, C^k\}$ for all the modalities.
**Initialize:** Divide $X$ into $M$ clusters randomly.
**repeat**
  **for** For every $x \in X$ **do**
    **Draw** $x$ from its current cluster.
    Calculate all the merge costs $\Delta\mathcal{L}$ based on Eq. (10).
    **Merge** $x$ into $t$ that $t = \arg\min_{t\in T}\Delta\mathcal{L}$.
  **end for**
**until** $p(t|x)$ is not changed

Substituting the Eq. (11) into Eq. (12), we can get that

$$\begin{aligned}&\Delta I_i^{private} = I(T^{bef};Y^i) - I(T^{aft};Y^i) = \\ &p(x)\sum_{y^i\in Y^i}p(y^i|x)\log\frac{p(y^i|x)}{p(y^i)} + p(t)\sum_{y^i\in Y^i}p(y^i|t)\log\frac{p(y^i|t)}{p(y^i)} - \\ &\sum_{y^i\in Y^i}p(x)p(y^i|x)\log\frac{p(y^i|\tilde{t})}{p(y^i)} - \sum_{y^i\in Y^i}p(t)p(y^i|t)\log\frac{p(y^i|\tilde{t})}{p(y^i)} = \\ &p(x)\sum_{y^i\in Y^i}p(y^i|x)\log\frac{p(y^i|x)}{p(y^i|\tilde{t})} + p(t)\sum_{y^i\in Y^i}p(y^i|t)\log\frac{p(y^i|t)}{p(y^i|\tilde{t})} \\ &= [p(x)+p(t)]\cdot JS_{\prod}[p(Y^i|x), p(Y^i|t)],\end{aligned} \tag{13}$$

where $\prod = \{\pi_1, \pi_2\} = \{\frac{p(x)}{p(x)+p(t)}, \frac{p(t)}{p(x)+p(t)}\}$, $JS_{\prod}$ is the *Jensen-Shannon* divergence [37]. Similarly, we can obtain the $\Delta I^{common}$ as follow

$$\begin{aligned}\Delta I^{common} &= I(T^{\text{bef}};\tilde{Y}) - I(T^{\text{aft}};\tilde{Y}) = \\ &[p(x)+p(t)]\cdot JS_{\Pi}[p(\tilde{Y}|x), p(\tilde{Y}|t)].\end{aligned} \tag{14}$$

As for the $\Delta I_i^{clustering}$, we can get the $I(T^{bef};C^i)$ and $I(T^{aft};C^i)$ according to Eq. (7). Now, the overall merge cost can be calculated. Next, we present the SPIB algorithm as in Algorithm 1.

### *E. THEORETICAL ANALYSIS*

#### 1) Convergence Analysis

**Theorem 1.** *The objective function of SPIB method is able to converge to a stable solution.*

*Proof.* We first prove that the value of Eq. (9) increases in every draw and merge iteration. We use $\mathcal{L}^{old}$ to indicate the value of Eq. (9) before $x$ is drawn from its current cluster $t^{old}$, and use $\mathcal{L}^{bef}$ and $\mathcal{L}^{aft}$ to indicate the value of Eq. (9) before and after $x$ is merged into some new cluster $t^{new}$. Thus, the merge process has two situations:

1) $t^{old} = t^{new}$. The value of Eq. (9) is no changed since the $\{x\}$ is merged into its original cluster in this situation, i.e., $\mathcal{L}^{bef} = \mathcal{L}^{aft}$.

2) $t^{old} \neq t^{new}$. Since $t^{new}$ satisfies that $t^{new} = \arg\min \Delta\mathcal{L} = \mathcal{L}^{bef} - \mathcal{L}^{aft}$, thus, the merge cost $\Delta\mathcal{L}(x, t^{new})$ must be smaller than $\Delta\mathcal{L}(x, t^{bef})$, i.e., $\Delta\mathcal{L}(x, t^{new}) < \Delta\mathcal{L}(x, t^{bef})$. Obviously, $\Delta\mathcal{L}(x, t^{new}) = \mathcal{L}^{bef} - \mathcal{L}^{old}$ and $\Delta\mathcal{L}(x, t^{new}) = \mathcal{L}^{bef} - \mathcal{L}^{aft}$. Thus, we can get $\mathcal{L}^{bef} \geq \mathcal{L}^{aft}$

Next, we prove the Eq. (9) is upper-bounded. Since $T$ is a compressed representation of $X$, there must have $I(T; Y^i) \leq I(X; Y^i)$ and $I(T; \tilde{Y}) \leq I(X; \tilde{Y})$. Moreover, suppose there is a true partition $C$ for the source dataset $X$, we can obtain $I(T; C^i) < I(T; C)$. Thus, the upper bound of Eq. (9) is $\sum_{i=1}^{k} I(X; Y^i) + \lambda \cdot [I(X; \tilde{Y}) + \sum_{i=1}^{k} I(T; C)]$. In summary, we can prove that the objective function of the SPIB method is able to converge to a stable solution. □

#### 2) Complexity Analysis

The Step 4 in Algorithm 1 is to construct hybrid words space, and its time complexity is $O(|\tilde{Y}|^2)$. The time complexity of step 5 is $O(n \log n)$ since we adopt the IB algorithm to build the basic clusterings for all modalities. The random partition in step 6 is $O(1)$. In the main loop of the SPIB algorithm, the time complexity of merge cost calculation in step 10 is $O(|X|(|Y^1| + |Y^2| + \cdots + |Y^k| + |\tilde{Y}|))$. Thus, the overall time complexity of the SPIB algorithm is $O(M|X|(|Y^1| + |Y^2| + \cdots + |Y^k| + |\tilde{Y}|))$, where $M$ is the number of clusters.

## IV. SEMANTIC EXTENSION FOR SOCIAL IMAGES

The proposed SPIB method can effectively cope with the cross-modal data when every modality has rich data information. With the increasing popularity of social media, large amounts of social images are being generated and collected everyday [19]–[21]. The social images include two modalities: image and user tags. In the applications of social media, most users usually upload images with very simple words or tags, which leads to rare textual information in social images [22]. Aiming at this problem, we propose a novel semantic extension model based on Gloss vector [40], in which the structured knowledge graph in the form of WordNet [41] is adopted as a large corpus. Then, we design a novel semantic similarity measure for social images.

### A. SEMANTIC EXTENSION BASED ON GLOSS VECTOR

Gloss vector is a semantic relatedness measurement algorithm, which bases on the assumption that one word can be characterized by its context. In this section, we employ the large corpus WordNet to obtain a word space with strong generalization ability. Once the word space is determined, we can generate a word vector for the word $w$ by following steps:

1) Initialize a fixed-order all zero vector $w$;

2) Query the interpretation of the word $w$ in WordNet, and extract the keywords in the interpretation. Then, we add the keywords to the corresponding position in the vector $w$ by adding 1;

TABLE 1: The details of the 8 cross-modal datasets

| Datasets | # Clusters | # Modalities | # Instances |
|---|---|---|---|
| Wikipedia | 10 | 2 | 2866 |
| Pascal Sentence | 20 | 2 | 1000 |
| Pascal VOC 07 | 20 | 2 | 9963 |
| X-Media | 20 | 2 | 5000 |
| NUS-WIDE | 6 | 2 | 2969 |
| IAPR TC-12 | 6 | 2 | 3095 |
| Reuters Multilingual | 6 | 5 | 15000 |
| HMDB | 51 | 3 | 6849 |

3) Repeat the step 2) for all the keywords.

Now, we can obtain the word vector for the word $w$ based on WordNet. The above process not only obtains the interpretation of a word, but also characterizes the second-order meaning of the keyword, thus, it can capture complete semantic information of the social images. And the similarity between two words can be calculated by the cosine similarity between the word vectors.

### B. SEMANTIC SIMILARITY MEASUREMENT

The semantic similarity between two social images can be computed according to the word similarities of user tags. If two words $w_1$ and $w_2$ are noun, we can employ the concept distance $C(w_1, w_2)$ between two words in WordNet, which is defined as follow

$$C(w_1, w_2) = -\log \frac{length(w_1, w_2)}{2depth} \tag{15}$$

where $length(w_1, w_2)$ is the number of nodes between $w_1$ and $w_2$ in WordNet. $depth$ is the maximum number of layers in WordNet, and we set it as 16.

When there at least one word is not noun, the cosine similarity $G(w_1, w_2)$ between the word vectors by the Gloss vector can be seen as the word similarity. Thus, the semantic similarity $S(w_1, w_2)$ between any two words can be calculated as follow

$$S(w_1, w_2) = \begin{cases} \dfrac{C(w_1, w_2) + G(w_1, w_2)}{2}, (w_1, w_2) \in noun \\ G(w_1, w_2), otherwise. \end{cases} \tag{16}$$

Now, the semantic similarity between two social images $I_1$ and $I_2$ can be calculated by their corresponding tags $I_1 = \{p_1, \cdots, p_r\}$ and $I_2 = \{q_1, \cdots, q_s\}$. Thus, the semantic similarity between $I_1$ and $I_2$ can be calculated as follow

$$Simi(I_1, I_2) = \frac{\sum_{i=1}^{r} S(p_i, q) + \sum_{j=1}^{s} S(p, q_j)}{2} \tag{17}$$

where $p$ and $q$ is the words in social image $I_1$ and $I_2$. Thus, the proposed SPIB method can cope with the social images after the semantic similarity matrix is constructed by the proposed semantic extension model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



TABLE 2: The AC (%) comparison of SPIB with all the baselines.

| Data sets | Single-modal clustering | | Concatenate-IB | Cross-modal clustering | | | | | SPIB |
|---|---|---|---|---|---|---|---|---|---|
| | $k$-means | IB | | CCA | CIB | CSPA | PTGP | LWEA | |
| Wikipedia | 45.22 ± 2.4 | 58.27 ± 4.8 | 28.58 ± 0.9 | 59.67 | 62.53 ± 2.6 | 60.09 ± 5.1 | 62.07 ± 1.0 | 62.86 ± 0.8 | **65.50 ± 2.3** |
| Pascal Sentence | 52.93 ± 3.8 | 54.10 ± 2.9 | 41.75 ± 2.8 | 55.20 | 57.24 ± 1.8 | 55.15 ± 2.0 | 57.83 ± 0.8 | 59.67 ± 0.7 | **63.02 ± 1.5** |
| Pascal VOC 07 | 64.99 ± 3.5 | 69.53 ± 3.2 | 62.77 ± 2.3 | 69.09 | 72.61 ± 2.1 | 69.84 ± 5.3 | 72.89 ± 0.9 | 73.64 ± 0.8 | **76.33 ± 1.2** |
| X-Media | 20.55 ± 0.6 | 21.16 ± 0.5 | 21.31 ± 0.5 | 23.58 | 22.14 ± 0.8 | 18.75 ± 0.7 | 24.31 ± 1.1 | 21.78 ± 0.6 | **25.16 ± 1.9** |
| NUS-WIDE | 46.11 ± 0.6 | 56.19 ± 0.5 | 57.61 ± 0.2 | 58.34 | 59.76 ± 1.1 | 59.26 ± 1.4 | 58.22 ± 1.8 | 60.21 ± 0.4 | **62.17 ± 2.2** |
| IAPR TC-12 | 68.22 ± 0.5 | 80.83 ± 0.1 | 81.02 ± 0.4 | 82.69 | 2.13 ± 1.5 | 81.63 ± 1.1 | 79.05 ± 1.3 | 83.62 ± 1.0 | **86.31 ± 1.5** |
| Reuters Multilingual | 53.16 ± 1.4 | 53.12 ± 3.1 | 53.43 ± 3.4 | 50.93 | 57.46 ± 1.3 | 59.92 ± 0.2 | 56.61 ± 3.6 | 55.28 ± 4.1 | **60.59 ± 3.6** |
| HMDB | 18.46 ± 0.9 | 22.31 ± 2.2 | 23.35 ± 1.3 | 25.34 | 27.20 ± 0.6 | 20.82 ± 0.5 | 25.75 ± 0.9 | 26.63 ± 1.4 | **29.42 ± 0.8** |
| Average | 42.21 | 51.94 | 46.23 | 53.11 | 54.88 | 53.18 | 54.59 | 55.46 | **58.56** |

TABLE 3: The NMI (%) comparison of SPIB with all the baselines.

| Data sets | Single-modal clustering | | Concatenate-IB | Cross-modal clustering | | | | | SPIB |
|---|---|---|---|---|---|---|---|---|---|
| | $k$-means | IB | | CCA | CIB | CSPA | PTGP | LWEA | |
| Wikipedia | 45.50 ± 1.7 | 51.68 ± 1.3 | 14.57 ± 0.4 | 52.32 | 53.38 ± 1.5 | 48.12 ± 2.3 | 52.98 ± 0.2 | 53.01 ± 0.3 | **54.97 ± 1.6** |
| Pascal Sentence | 59.88 ± 1.9 | 59.99 ± 1.6 | 47.88 ± 1.8 | 60.89 | 60.21 ± 1.1 | 61.99 ± 1.4 | 62.12 ± 0.4 | 63.75 ± 0.3 | **66.01 ± 1.8** |
| Pascal VOC 07 | 65.16 ± 1.5 | 65.38 ± 1.2 | 65.29 ± 1.0 | 66.28 | 69.74 ± 2.3 | 71.05 ± 2.4 | 70.41 ± 0.5 | 71.92 ± 0.2 | **75.21 ± 1.5** |
| X-Media | 20.12 ± 0.5 | 21.24 ± 0.2 | 21.23 ± 0.2 | 23.18 | 23.17 ± 2.7 | 21.45 ± 0.8 | 23.94 ± 0.4 | 20.03 ± 0.3 | **26.54 ± 2.2** |
| NUS-WIDE | 39.42 ± 0.4 | 48.40 ± 0.5 | 42.84 ± 0.2 | 49.72 | 51.29 ± 1.0 | 50.38 ± 1.5 | 50.28 ± 1.1 | 52.61 ± 0.9 | **54.28 ± 1.3** |
| IAPR TC-12 | 64.83 ± 0.2 | 64.52 ± 0.1 | 62.03 ± 0.5 | 64.18 | 66.87 ± 1.8 | 62.59 ± 0.9 | 66.87 ± 1.7 | 67.32 ± 1.4 | **70.54 ± 2.3** |
| Reuters Multilingual | 33.94 ± 1.2 | 43.29 ± 3.1 | 44.33 ± 3.4 | 46.77 | 50.28 ± 1.7 | 48.80 ± 0.1 | 47.35 ± 4.9 | 46.79 ± 4.0 | **51.79 ± 2.8** |
| HMDB | 26.84 ± 1.0 | 33.74 ± 2.2 | 34.15 ± 2.0 | 36.14 | 37.46 ± 0.9 | 37.83 ± 0.6 | 37.25 ± 0.6 | 35.41 ± 0.8 | **41.13 ± 1.1** |
| Average | 44.84 | 48.53 | 41.54 | 49.94 | 51.43 | 50.28 | 51.4 | 51.36 | **55.06** |

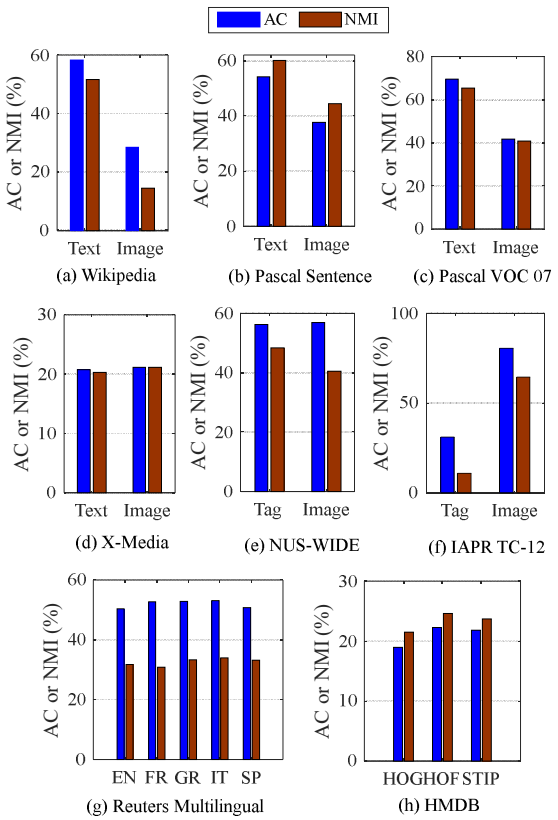


FIGURE 5: The clustering results of IB on single modality.

## V. EXPERIMENTS

### A. DATASETS AND EVALUATION METRICS

In this study, we employ 4 types of cross-modal datasets to demonstrate the effectiveness of the proposed SPIB method. The details of the datasets are shown in Table 1.

**1) Cross-media dataset**. It contains 4 cross-modal datasets: Wikipedia [42], Pascal Sentence [43], Pascal VOC 07 [44] and X-Media [45]. They contains image-text pairs from various categories, and each image accompanies a text document. For Wikipedia, Pascal Sentence and X-media, we extract the popular 128 dimension SIFT [46] features to represent the images. For text representation, we first obtain the feature vector based on 500 tokens (with stop words removed), and then the LDA model is used to reduce the dimension to 100 dimensional probability vectors. For the Pascal VOC 07 dataset, we employ 776 dimensional visual feature, which contains a 200 dimensional SIFT BoVW feature, a 512-D GIST feature and a 64 dimensional HSV feature. For the text representation, we employ a 798 dimensional tag ranking feature provided by [47].

**2) Social images dataset**. It contains two popular social image datasets: NUS-WIDE [48] and IAPR TC-12 [49]. The NUS-WIDE is a social image dataset from national university of singapore, which consists of social images and its user tags. The NUS-WIDE includes 6696 tags and the number of tags in each image are from 3 to 12. The IAPR TC-12 is a image annotation dataset from CLEF, in which each image contains a short descriptive text message.

**3) Reuters Multilingual dataset**. The Reuters Multilingual dataset [1] contains 6 categories (i.e., C15, CCAT, E21, ECAT, GCAT and M11) of news reported by 5 different languages (i.e., Spanish, Italian, German, French and English). All documents are represented by using the 1000 dimensional BoW feature in this study.

**4) Multiple heterogenous feature dataset**. The HMDB [50] dataset consists of 51 categories of 6849 human action video sequences, mainly recorded from movie clips,

web vides, etc. In this study, we extract three heterogeneous features: histogram of oriented gradient (HOG), histogram of optical flow (HOF) and space-time interest points (STIP). Then, a 1000-dimensional BoVW representation for all feature representation is built separately.

We employ the most popular evaluation metrics, i.e., normalized mutual information (NMI) and clustering accuracy (AC) [25], to measure the experimental results.

### B. BASELINES

We compare the proposed SPIB with the 3 types of baseline algorithms:

**1) Single-modal clustering algorithms**. $k$-means and IB [31] are two typical and effective clustering algorithms. We report the best clustering results for all modalities in the tables and figures in this study.

**2) Concatenate-IB algorithm**. We concatenate the features of multiple modalities, and use the IB method to cluster the data instances based on mutual information.

**3) Cross-modal clustering algorithms**. In this study, 4 state-of-the-art cross-modal clustering algorithms are adopted, which are canonical correlation analysis (CCA) [12], consensus information bottleneck (CIB) [26], cluster-based similarity partitioning algorithm (CSPA) [25], probability trajectory graph partitioning (PTGP) [29] and locally weighted evidence accumulation (LWGP) [30]. The CCA algorithm can learn the shared information of two modalities. Thus, for the cross-modal datasets with more than 2 modalities, we select two modalities with best performances as the input of CCA algorithm. The CSPA, PTGA and LWGP are the state-of-the-art consensus clustering, which adopt co-association matrix, probability trajectory and evidence accumulation to integrate the basic clusterings of all modalities.

To fairly compare all the baselines, we adopt the source codes provided by the corresponding authors with the default or optimal parameter settings of their original papers.

### C. EXPERIMENTAL RESULTS AND ANALYSIS

#### 1) The Performance of Each Modality

In order to verify the representation ability of different modalities of the cross-modal datasets, we perform the original IB algorithm on the different modalities of cross-modal dataset. As shown in Fig. 5, we can get the following observations.

1) For Wikipedia, Pascal sentence and Pascal VOC 07 datasets, the performance of IB algorithm on the text modality is always better than the image modality. This phenomenon shows that it is difficult to obtain better clustering results only by image modality in cross-modal image clustering tasks. The rich semantic information can lead to more accurate clustering results.

2) For the social image datasets IAPR TC-12 and NUS, we utilize the semantic extension method based on Gloss vector to construct the semantic similarity matrix. As shown in Fig 5, the IB algorithm gets comparable clustering performance on the constructed semantic similarity matrix compared with the image modality on NUS-WIDE dataset. This verifies the effectiveness of the proposed semantic extension method. For IAPR TC-12 dataset, the performance of the IB algorithm on the constructed semantic similarity matrix is not better than the image modality. This is mainly because the semantic information in IAPR TC-12 is a sentence with many function words, and the notional words are too rare to characterize its semantic information.

3) The IB algorithm obtains similar clustering results on the heterogeneous feature spaces of Reuters multi-lingual dataset and the HMDB human action video dataset. This phenomenon shows that data instances can often be described from different feature representations, which reflect the different intrinsic characteristics of the cross-modal dataset. Therefore, it is wise to improve the quality of final clustering results by effectively organizing the complementary effect of multiple modalities.

#### 2) The Comparison with Cross-modal Clustering Methods

Table 2 and Table 3 show the comparison results of the proposed SPIB with all the other baselines. From these two tables, we can get the following observations.

1) Directly concatenating multiple modalities cannot always improve the clustering performance. For instance, the concatenate-IB algorithm obtains lower AC and NMI values than the original IB on the Pascal Sentence, and Pascal VOC 07 datasets. It is stated that simply connecting the multiple modalities cannot stably improves the clustering quality.

2) The performances of cross-modal clustering algorithms are much better than the best results of single-modal clustering algorithms. For instance, the average AC and NMI values of the four cross-modal clustering algorithms are all better than that of single-modal clustering algorithms.

3) Compared with other single-modal and cross-modal clustering baseline algorithms, the proposed SPIB obtains significant improvements on all the cross-modal datasets used in this paper. This phenomenon verifies the effectiveness of the proposed SPIB method on the task of cross-modal clustering.

### D. THE EXPLORATION OF IMPACT FACTORS

#### 1) Parameter Analysis

The proposed SPIB method utilizes the parameter $\lambda$ to control the balance between shared information and private information. Thus, we conduct experiment to evaluate the impact of $\lambda$ on the performance of the SPIB algorithm. Specifically, we vary the value of $\lambda$ from the space $\{10, 20, \cdots, 110\}$.

It can be seen from Fig. 6 that the SPIB algorithm obtains a low AC value when the $\lambda$ is small. Then, the AC values of the SPIB algorithm gradually improve as the value of $\lambda$ increases. This is mainly because the complementary effect of the shared and private information begins to take effect. And the AC and NMI values of the SPIB algorithm reported in this paper are the clustering results when $\lambda = 60$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2904554, IEEE Access



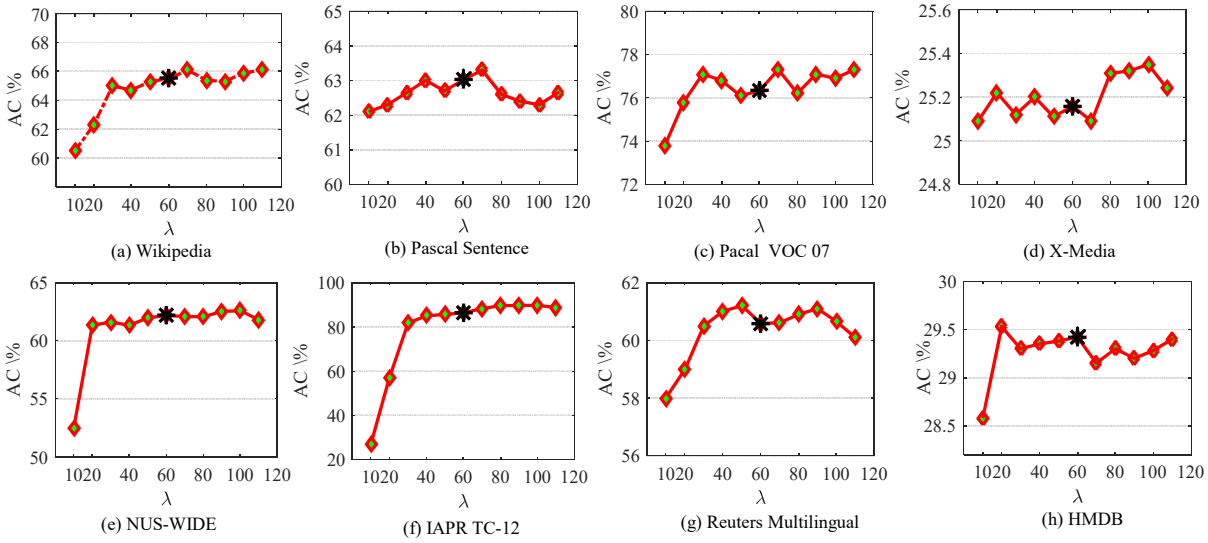


FIGURE 6: The performance of SPIB algorithm with various $\lambda$.

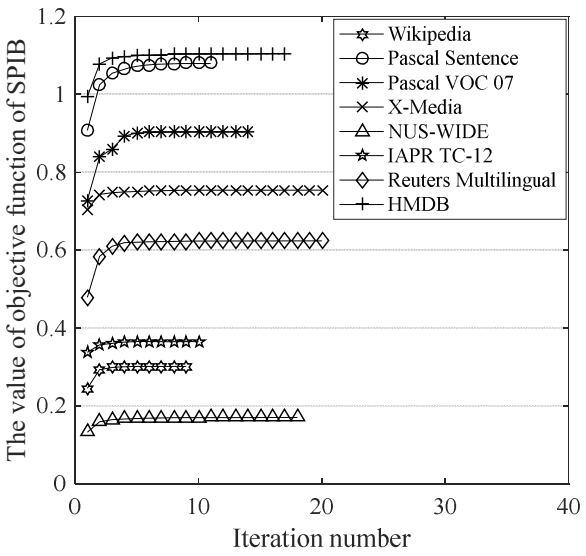


FIGURE 7: The convergence of SPIB algorithm.

### 2) Convergence Analysis

The objective function (9) of the SPIB algorithm can only converge to a local optimal solution, thus, it is necessary to empirically analyze its convergence. Fig. 7 shows the value of Eq. (9) at each iteration of the SPIB algorithm. It can be seen from this figure that the value of Eq. (9) rises rapidly at first, then the amplitude of the rise tends to be flat, and finally the optimal value of the objective function is obtained.

### 3) Ablation Test

The proposed SPIB algorithm aims to discover more reasonable cluster structure by simultaneously considering the shared information of multiple modalities and the private information of individual modality, which are constructed by the novel hybrid words model and consensus clustering model, respectively. Thus, we conduct experiment to verify the effectiveness of each individual model in this section.

TABLE 4: The AC value of SPIB on individual modality. In this table, we use CC and Hybrid to indicate the consensus clustering model and hybrid words model, respectively.

| Datasets | IB | CC | Hybrid | SPIB |
|---|---|---|---|---|
| Wikipedia | $58.27 \pm 4.8$ | $62.18 \pm 4.3$ | $61.99 \pm 1.7$ | $\mathbf{65.50 \pm 2.3}$ |
| Pascal Sentence | $54.10 \pm 2.9$ | $56.74 \pm 2.8$ | $55.41 \pm 3.7$ | $\mathbf{63.02 \pm 1.5}$ |
| Pascal VOC 07 | $69.53 \pm 3.2$ | $70.23 \pm 2.3$ | $72.38 \pm 1.6$ | $\mathbf{76.33 \pm 1.2}$ |
| X-Media | $21.16 \pm 0.5$ | $20.58 \pm 0.4$ | $23.76 \pm 0.8$ | $\mathbf{25.16 \pm 1.9}$ |
| NUS-WIDE | $56.19 \pm 0.5$ | $58.31 \pm 0.9$ | $59.91 \pm 1.5$ | $\mathbf{62.17 \pm 2.2}$ |
| IAPR TC-12 | $80.83 \pm 0.3$ | $83.42 \pm 1.6$ | $84.38 \pm 1.8$ | $\mathbf{86.31 \pm 1.5}$ |
| Reuters | $53.12 \pm 3.1$ | $57.79 \pm 3.4$ | $58.09 \pm 2.2$ | $\mathbf{60.59 \pm 3.6}$ |
| HMDB | $22.31 \pm 2.2$ | $23.32 \pm 2.4$ | $26.14 \pm 1.3$ | $\mathbf{29.42 \pm 0.8}$ |
| Average | 51.94 | 50.07 | 55.26 | **58.56** |

Table 4 shows the clustering results of the proposed SPIB method when considering individual hybrid words model and consensus clustering model. From Table 4, we can observe: 1) The clustering results of the SPIB algorithm with single hybrid words model and consensus clustering model are better than the optimal results of the IB algorithm on single modality. 2) The average results of the SPIB algorithm on consensus clustering model are slight better than the hybrid words model. 3) When simultaneously considering both the hybrid words model and the consensus clustering model, the clustering results of the SPIB algorithm is further improved. This phenomenon verifies the effectiveness of the hybrid words model and consensus clustering model.

## VI. CONCLUSION

In this study, we propose a novel shared-private information bottleneck method for cross-modal clustering. Firstly, two novel shared information construction models are proposed to build the shared information of different modalities. Then,

the shared information of multiple modalities and the private information of individual modalities are maximally preserved through a unified information maximization objective function. Finally, a sequential "draw-and-merge" procedure is proposed to optimize the objective function of SPIB. Besides, to solve the lack of social tags modality in cross-modal social images, we also investigate the use of structured prior knowledge in the form of knowledge graph to enrich the information in semantic modality, and design a novel semantic similarity measurement for social images. Experimental results on cross-modal clustering tasks demonstrate that our method outperforms the state-of-the-art approaches.

## REFERENCES

[1] A. Kumar and H. D. III, "A co-training approach for multi-view spectral clustering," in Proceedings of the International Conference on Machine Learning (ICML), 2011, pp. 393–400.

[2] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1977–1984.

[3] J. Zhang and Y. Peng, "Query-adaptive image retrieval by deep-weighted hashing," IEEE Transactions on Multimedia (TMM), vol. 20, no. 9, pp. 2400–2414, 2018.

[4] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," IEEE Transactions on Image Processing (TIP), vol. 28, no. 4, pp. 1602–1612, 2019.

[5] Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval," IEEE Transactions on Image Processing (TIP), vol. 26, no. 3, pp. 1393–1404, 2017.

[6] A. Wu and Y. Han, "Multi-modal circulant fusion for video-to-language and backward," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 1029–1035.

[7] H. Liu, G. Xiao, Y. Tan, and C. Ouyang, "Multi-source remote sensing image registration based on contourlet transform and multiple feature fusion," International Journal of Automation and Computing, pp. 1–14, 2018.

[8] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang, "LBMCH: learning bridging mapping for cross-modal hashing," in Proceedings of the 38th International ACM Conference on Research and Development in Information Retrieval (SIGIR), 2015, pp. 999–1002.

[9] X. Zhang, Q. Yu, and H. Yu, "Physics inspired methods for crowd video surveillance and analysis: A survey," IEEE Access, vol. 6, pp. 66 816–66 830, 2018.

[10] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[11] Y. Wu, Y. Ye, C. Zhao, and Z. Shi, "Collective density clustering for coherent motion detection," IEEE Transactions on Multimedia (TMM), vol. 20, no. 6, pp. 1418–1431, 2018.

[12] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 129–136.

[13] R. Memisevic, L. Sigal, and D. J. Fleet, "Shared kernel information embedding for discriminative inference," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 34, no. 4, pp. 778–790, 2012.

[14] C. Jin, W. Mao, R. Zhang, Y. Zhang, and X. Xue, "Cross-modal image clustering via canonical correlation analysis." in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2015, pp. 151–159.

[15] C. H. Ek, P. H. Torr, and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," in International Workshop on Machine Learning for Multimodal Interaction (MLMI), 2007, pp. 132–143.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research (JMLR), vol. 3, no. Jan, pp. 993–1022, 2003.

[17] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in Proceedings of the SIAM International Conference on Data Mining (ICDM), 2013, pp. 252–260.

[18] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data." in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp. 2598–2604.

[19] X. Chen, A. Ritter, A. Gupta, and T. Mitchell, "Sense discovery via co-clustering on images and text," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5298–5306.

[20] S. Hong, J. Choi, J. Feyereisl, B. Han, and L. S. Davis, "Joint image clustering and labeling by matrix factorization," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 7, pp. 1411–1424, 2016.

[21] X. Xu, L. He, H. Lu, A. Shimada, and R.-I. Taniguchi, "Non-linear matrix completion for social image tagging," IEEE Access, vol. 5, pp. 6688–6696, 2017.

[22] J. Wang, X. Zhu, and S. Gong, "Video semantic clustering with sparse and incomplete tags." in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2016, pp. 3618–3624.

[23] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, 2001, pp. 408–415.

[24] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," Journal of Machine Learning Research (JMLR), vol. 3, pp. 1107–1135, 2003.

[25] A. Strehl and J. Ghosh, "Cluster ensembles–a knowledge reuse framework for combining multiple partitions," Journal of Machine Learning Research (JMLR), vol. 3, no. Dec, pp. 583–617, 2002.

[26] X. Yan, Y. Ye, and X. Qiu, "Unsupervised human action categorization with consensus information bottleneck method," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 2245–2251.

[27] I. Khan and Z. Luo, "Nonnegative matrix factorization based consensus for clusterings with a variable number of clusters," IEEE Access, vol. 6, pp. 73 158–73 169, 2018.

[28] S. Nejatian, H. Parvin, and E. Faraji, "Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification," Neurocomputing, vol. 276, pp. 55–66, 2018.

[29] D. Huang, J.-H. Lai, and C.-D. Wang, "Robust ensemble clustering using probability trajectories," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 28, no. 5, pp. 1312–1326, 2016.

[30] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," IEEE Transactions on Cybernetics (TCYB), vol. 48, no. 5, pp. 1460–1473, 2018.

[31] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in Proceedings of the Annual Allerton Conference on Communnication, Control and Computing, 1999, pp. 368–377.

[32] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR), 2000, pp. 208–215.

[33] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 36, no. 8, pp. 1559–1572, 2014.

[34] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1496–1505.

[35] X. Yan, Y. Ye, and Z. Lou, "Unsupervised video categorization based on multivariate information bottleneck method," Knowledge-Based Systems (KBS), vol. 84, no. C, pp. 34–45, 2015.

[36] X. Yan, S. Hu, and Y. Ye, "Multi-task clustering of human actions by sharing information," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6401–6409.

[37] T. M. Cover and J. A. Thomas, Elements of information theory (2. ed.). Wiley, 2006.

[38] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proceedings of the European Conference on Machine Learning (ECML), 1998, pp. 137–142.

[39] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[40] S. Patwardhan and T. Pedersen, "Using wordnet-based context vectors to estimate the semantic relatedness of concepts," in Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together, 2006.

[41] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13–47, 2006.

[42] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in Proceedings of the ACM international Conference on Multimedia (ACM'MM), 2010, pp. 251–260.

[43] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp. 139–147.

[44] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision (IJCV), vol. 88, no. 2, pp. 303–338, 2010.

[45] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 26, no. 3, pp. 583–596, 2016.

[46] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[47] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in Proceedings of the British Machine Vision Conference (BMVC), 2010, pp. 1–12.

[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), 2009, pp. 8–10.

[49] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in Proceedings of the International Conference on Language Resources and Evaluation (ICLRE), 2006, pp. 21–32.

[50] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2556–2563.