CMIB: Unsupervised Image Object Categorization in Multiple Visual Contexts

Xiaoqiang Yan, *Member, IEEE,* Yangdong Ye, *Member, IEEE,* Xueying Qiu, Milos Manic, *Senior Member, IEEE,* Hui Yu, *Senior Member, IEEE,*

***Abstract*—Object categorization in images is fundamental to various industrial areas, such as automated visual inspection, fast image retrieval and intelligent surveillance. Most existing methods treat visual features (e.g., scale-invariant feature transform, SIFT) as content information of the objects, while regarding image tags as their contextual information. However, the image tags can hardly be acquired in completely unsupervised settings, especially when the image volume is too large to be marked. In this work, we propose a novel contextual multivariate information bottleneck (CMIB) method to conduct unsupervised image object categorization in multiple visual contexts. Unlike using manual contexts, the CMIB method first automatically generates a set of high-level basic clusterings by multiple global features, which are unprecedentedly defined as visual contexts since they can provide overall information about the target images. Then, the idea of the data compression procedure for object category discovery is proposed, in which the content and multiple visual contexts are maximally preserved through a "bottleneck". Specifically, two Bayesian networks are initially built to characterize the relationship between data compression and information preservation. Finally, a novel sequential information-theoretic optimization is proposed to ensure the convergence of the CMIB objective function. Experimental results on seven real-world benchmark image datasets demonstrate that the CMIB method achieves better performance than the state-of-the-art baselines.**

***Index Terms*—Object category discovery, visual contexts, information bottleneck, mutual information, Bayesian networks.**

I. Introduction

OBJECT categorization in images has been an active and fundamental research topic, and a promising image clustering algorithm lays a good foundation for various industrial areas, such as automated visual inspection [1], fast image retrieval [2], [3] and intelligent surveillance [4], [5]. Recently, contextual information, a type of available and complementary information that provides rich positive details for target data, has been used to enhance the accuracy of object categorization models. In the task of object recognition and categorization of images, a large number of studies [6], [7], [8], [9], [10], [11], [12], [13] have also shown the validity of contextual information.

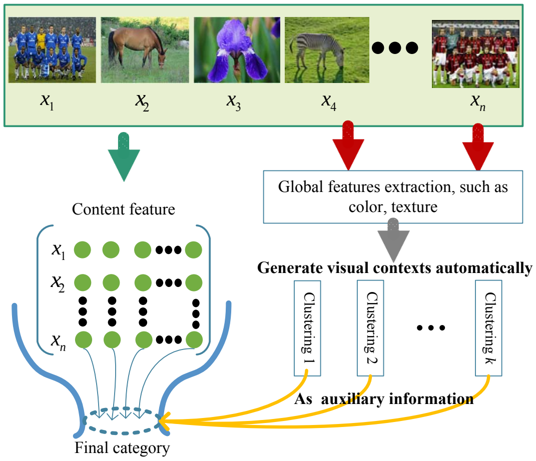


Fig. 1. The pipeline of unsupervised image object categorization in multiple visual contexts based on the proposed CMIB. First, CMIB adopts content feature to characterize the local information of the target object, while automatically generating a set of high-level basic clusterings by multiple global features, which are unprecedentedly defined as visual contexts in this study. Then, the object category discovery in unlabeled images is creatively formulated as a data compression procedure, in which the content and multiple visual contexts are maximally preserved through a "bottleneck". In other words, automatically constructed visual contexts can be utilized as auxiliary information to improve the quality of image clustering based on content feature. Note that, it is still unsupervised since the visual contexts come from the unlabeled image data.

Generally, the most popular strategy is to use the artificial marks of the target images as their contexts, such as user tags, captions and concurrent texts of the target images. However, existing methods always assume that the labeled training data are typically available for both the images and its contexts [6], [7], [8], [9], [10], [13], [14], which allows direct inference of the relationship between the image object categories and their contexts. Obviously, the ground-truth label and most types of contexts are often scarce and precious since they require a considerable amount of manpower and material resources, especially when the size of the image data is too large to be marked. This motivates us to conduct the study on unsupervised image object categorization by automatically constructing contexts from the unlabeled image data.

Recently, several types of contexts based on multiple feature representations have been proposed for the task of unsupervised object categorization in unlabeled images. For instance, [11], [12] proposed a context-aware clustering algorithm, in which the spatial neighbors of multiple primitive features are treated as their spatial contexts. [15] introduced a dual

This work is supported by the National Natural Science Foundation of China under grant No. 61772475, No. 61906172, the National Key Research and Development Program of China under grant No. 2018YFB1201403 and the EPSRC through project 4D Facial Sensing and Modelling (EP/N025849/1). (*Corresponding author: Yangdong Ye*)

X. Yan, Y. Ye and X. Qiu are with the School of Information Engineering, Zhengzhou University, 450052, P.R. China (e-mail: iexqyan@zzu.edu.cn, ieydye@zzu.edu.cn, iexyqiu@gs.zzu.edu.cn)

M. Manic is with the School of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E2254 (e-mail: misko@ieee.org)

H. Yu is with the School of Creative Technologies, University of Portsmouth, PO1 2DJ, United Kingdom (email: hui.yu@port.ac.uk)

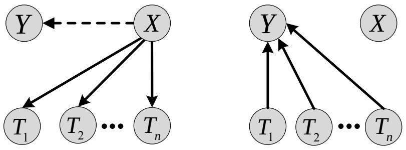
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2939278, IEEE Transactions on Industrial Informatics

assignment k-means algorithm for action clustering, which finds action categories by utilizing the scene features as the target action's contextual information, i.e., the scene that the target action takes place. As shown in the above methods, different possible feature representations are complementary to each other and can be naturally treated as the contexts of the target objects, such as spatial neighbors of multiple features [11], [12] and scene features [15]. However, existing methods always directly incorporate the content and contextual information, which may be problematic since the content and contexts are heterogeneous [16]. More importantly, the visual features are usually represented by several high dimensional descriptors, and dealing with them simultaneously always results in the curse of dimensionality.

In this work, a novel, general-purpose contextual multivariate information bottleneck (CMIB) is proposed to discover object categories in unlabeled images by devising multiple visual contexts. As shown in Fig. 1, the CMIB method adopts one content feature (such as SIFT) to characterize the local information of the target objects, while using global features to describe the object's contexts, such as global shape, color and texture. Instead of directly incorporating content and context features, CMIB method unprecedentedly defines the high-level clusterings automatically generated by multiple global features as multiple visual contexts, which naturally have the ability to leverage complementary information from heterogeneous raw content and context features. In particular, the automatically constructed visual contexts are group-level partitions of the target objects and can be utilized as auxiliary information to improve the quality of image clustering based on content feature. Note that it is still unsupervised since the visual contexts come from the unlabeled image data without any prior knowledge. Then, the task of unsupervised image object categorization in multiple visual contexts is formulated as an information maximization function, in which two novel Bayesian networks are built to characterize the relationships between the content and multiple visual contexts. Moreover, a novel sequential information-theoretic solution is proposed to ensure the convergence of the objective function of CMIB. The main contributions of this work are as follows:

- A novel CMIB is proposed for unsupervised object categorization in totally unlabeled images by devising multiple visual contexts. It is the first attempt to utilize the high-level clustering generated by global features as multiple visual contexts, which naturally have the ability to leverage heterogeneous features.
- The object category discovery in unlabeled images is creatively formulated as a data compression procedure through a "bottleneck", in which two Bayesian networks are built initially to characterize the relations between the data compression and information preservation.
- A novel sequential information-theoretic optimization is proposed to ensure the convergence of the objective function of CMIB. The proposed technique is completely unsupervised and outperforms the existing state-of-the-art baselines on several benchmark datasets.

This paper is structured as follows. Section II analyzes the related work about context-based clustering methods and the background of MIB. Section III formulates the proposed CMIB. In Section IV, we report and discuss the experimental results. Finally, Section V concludes the paper.



(a) $\mathcal{I}^{G_{in}}$ Data compression (b) $\mathcal{I}^{G_{out}}$ Information preservation

Fig. 2. Structural representation of G_{in} and G_{out} in the MIB framework. (a) In the data compression part, the solid arrows from X to $T_1, \cdots, T_n$ represent multiple compressed representations of X, while the dashed arrow between X and Y indicates a joint probability distribution $p(X, Y)$. (b) In the information preservation part, the solid arrows denote the information that should be maximized with respect to Y. $\mathcal{I}^{G_{in}}$ and $\mathcal{I}^{G_{out}}$ are the amount of information in these two networks.

II. Related work

A. Unsupervised Context-Based Methods

Recently, context-based clustering [11], [12], [17], [13], [15], [18] has been a valuable unsupervised learning topic in machine learning and its various industrial applications. However, all the existing context-based clustering approaches treat the target object's content and contextual information equivalently. In practice, the content and contextual information of the same object have their own structures, and their distributions are always heterogeneous. Thus, inappropriately integrating them will degrade the clustering performance.

Essentially, contextual information is a complement to the content information of an object. In this regard, it is pertinent to discuss multi-view and ensemble clustering methods. Both aim to improve the clustering performance by considering the complementary effect of multiple related components. Specifically, multi-view clustering methods [19], [20], [21], [22] aim to construct mappings, connections or agreements between multiple distinct views. In multi-view methods, the complementary views can be seen as the contexts of other views. In particular, low-level visual features are the most prevalent means to represent the different views of the objects in images. However, the dimensions of these visual features are always very high especially when dealing with them simultaneously. Moreover, how to control the balance of different views also remains a challenging task.

Different from multi-view clustering methods, ensemble clustering approaches [23], [24] refer to combining different clusterings of a given data collection into a single partition that is a better fit than existing clusterings. The information provided by auxiliary clusterings can also be seen as the context of the target data. However, the existing ensemble clustering methods yield the final partition without accessing the original feature representations of the images, which limits the final results in terms of the quality of existing base clusterings. In this study, we intend to discover the object category in unlabeled images by considering its content and visual contexts simultaneously.

B. Multivariate information bottleneck

The information bottleneck (IB) method [25] is one of the best among many clustering algorithms previously proposed, and concentrates on searching the cluster structure T of single source data X while only Y serves as a relevant variable. As a multivariate extension of IB, multivariate information bottleneck (MIB) [26], [27] uses the concept of *multi-information* to quantify the shared information between more than two variables, an example problem of multiple variables is shown in Fig. 2. Given a set of random variables $\mathbf{X} = \{X_1, \cdots, X_n\}$, the multi-information can be defined as

$$\mathcal{I}(\mathbf{X}) = D_{KL}[p(\mathbf{X})||p(X_1)\cdots p(X_n)]. \quad (1)$$

where D_{KL} indicates the Kullback-Leibler divergence [28].

In the MIB framework, the probabilistic model Bayesian network is adopted to characterize the relationships of multiple variables. Given a set of random variables $\mathbf{X} = \{X_1, \cdots, X_n\}$ and a set of latent variables $\mathbf{T} = \{T_1, \cdots, T_n\}$, a Bayesian network with graph G_{in} indicates the compressed relationship from $\mathbf{X}$ to $\mathbf{T}$. Another Bayesian network with graph G_{out} represents which conditional dependencies and independencies we want T to be able to generate. Both G_{in} and G_{out} are defined over $\mathbf{X} \bigcup \mathbf{T}$. Thus, the objective function of MIB can be defined as follows:

$$\mathcal{L}_{min}[p(\mathbf{T}|\mathbf{X})] = \mathcal{I}^{G_{out}}(\mathbf{X}, \mathbf{T}) - \beta^{-1} \cdot \mathcal{I}^{G_{in}}(\mathbf{X}, \mathbf{T}), \quad (2)$$

where $p(\mathbf{T}|\mathbf{X})$ is the mapping from $\mathbf{X}$ to $\mathbf{T}$, β strikes a balance between the data compression information preservation in G_{in} and G_{out}.The multi-information $\mathcal{I}^G$ with respect to a Bayesian network G defined over $\mathbf{X} \sim p(X)$ is computed as

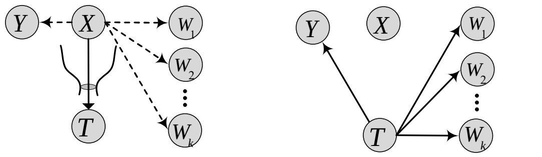
$$\mathcal{I}^G(\mathbf{X}) = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}), \quad (3)$$

where $\mathbf{Pa}^G_{X_i}$ are the parents of X_i in G, and $I(X_i; \mathbf{Pa}^G_{X_i})$ is the mutual information between X_i and $\mathbf{Pa}^G_{X_i}$.

MIB can manage multiple variables by utilizing the concept of multi-information. However, the original MIB method assumes that different information sources have completely homogenous data distributions. Obviously, this strict assumption is a major disadvantage of the original MIB since different data sources are often heterogeneous. In the proposed CMIB framework, to relieve the distinct gap between heterogeneous feature spaces, the high-level basic clusterings generated by the multiple global features are unprecedentedly defined as multiple visual contexts, which naturally have the ability to leverage complementary information from heterogeneous sources. To the best of our knowledge, this is the first work to define multiple visual contexts to relieve the distinct gap among heterogeneous features.

III. Contextual multivariate information bottleneck

Most existing object categorization methods treat visual features (e.g., scale-invariant feature transform, SIFT) as content information of the objects, while regarding artificial marks as the contextual information. However, artificial marks are not available in complete unsupervised settings, especially when the image volume is too large to be marked. In this paper, we propose a novel unsupervised object categorization method called CMIB, that can discover the object category in unlabeled images by simultaneously considering content feature and multiple visual contexts.



(a) $\mathcal{I}^{G_{in}}$ Data compression (b) $\mathcal{I}^{G_{out}}$ Information preservation

Fig. 3. The structural model of the CMIB method for unsupervised object category discovery in images. (a) In this part, the solid arrow from X to T means X is mapped into its compressed representation T. At the same time, the dashed arrows from X to Y and $W_1, W_2, \cdots, W_k$ indicate that the source data X has one content variable and multiple visual contexts. (b) In this part, the solid arrows specify the information contained in the content information Y and visual contexts $W_1, W_2, \cdots, W_k$ are maximally preserved.

A. Problem formulation

In this section, we explain the novel unsupervised object categorization method called CMIB, which discovers the object categories in unlabeled images by simultaneously considering one content feature and multiple visual contexts. First, we give the definition of the visual contexts.

Definition 1 (**Visual Contexts**). *Given an unlabeled image collection X and its k global feature representations, the visual contexts of the objects are the clusterings $W_1, W_2, \cdots, W_k$ constructed by the k global features.*

In this study, we utilize the source variable X, content variable Y, multiple visual contexts $W_1, W_2, \cdots, W_k$ and the final category (cluster) T to characterize the problem of CMIB. The source variable X denotes an unlabeled image collection with a set of samples $\{x_1, x_2, \cdots, x_n\}$. The content variable Y indicates one discriminative content feature (such as SIFT) of the image collection X. Correspondingly, we can construct the joint distribution $p(X, Y)$ between the source data and its content feature according to the prevalent bag-of-visual-words (BoVW) model [29], and the details can be found in Section IV-B. Then, any clustering algorithm with promising performance (e.g., IB algorithm in this study) can be utilized to construct multiple basic clusterings $W_1, W_2, \cdots, W_k$ according to the other k global features. For clarity, we first define the task of CMIB.

Definition 2 (**CMIB**). *Given an source variable X taking value from $\{x_1, x_2, \cdots, x_n\}$, there are one content variable Y that indicates the content feature and k visual contexts $W_1, W_2, \cdots, W_k$ that denote the multiple clusterings constructed by the k global features. The goal of CMIB is to discover the potential object categories $T = \{t_1, \cdots, t_M\}$ hidden in the unlabeled image collection X, where M is the number of categories. In other words, the task of CMIB is to find an optimal encoding scheme $p(t|x)$ from X to T, while* *maximally maintaining the information of the content variable and multiple visual contexts.*

B. Objective Function of CMIB

In this section, the objective function of the CMIB algorithm is presented. CMIB treats one local feature variable of the object as the content information Y, while exploiting the base clusterings $W_1, W_2, \cdots, W_k$ obtained by other global features as its visual contexts. As illustrated in Fig. 3, CMIB involves two parts: data compression and information preservation. We construct two Bayesian networks G_{in} and G_{out} to characterize the relationships between the variables in the two parts. In Bayesian network G_{in}, the dashed edges indicates that the source variable X has multiple relevant information. Specifically, the dashed edge $X \to Y$ means the relation defined by the joint distribution $p(X, Y)$. The dashed edge $X \to W_1, W_2, \cdots, W_k$ indicates that the source data X has multiple visual contexts $W_1, W_2, \cdots, W_k$. The solid edge $X \to T$ means that X will be compressed into its compressed representation T. In Bayesian network G_{out}, the solid edge $T \to Y$ reflects that T should capture the local content information Y about source data X. The solid edge $T \to W_1, \cdots, W_k$ reflects that T should capture all the visual context information $W_1, W_2, \cdots, W_k$. In other words, the network G_{in} represents the compressing relationship from X to T, while G_{out} expresses that T should simultaneously preserve the relevant information of content information Y and multiple visual contexts $W_1, W_2, \cdots, W_k$. Similar to [27], the multi-information in G_{in} and G_{out} can be defined as follows:

$$\mathcal{I}^{G_{in}} = I(X;T) + I(X;Y) + \textstyle\sum_{i=1}^{k} I(X;W_i), \tag{4}$$

$$\mathcal{I}^{G_{out}} = I(T;Y) + \textstyle\sum_{i=1}^{k} I(T;W_i), \tag{5}$$

where $I(X;T)$ is the term of mutual information [28] that measures how many bits are conveyed from the source variable X to its compressed representation T. $I(T;Y)$ measures how much information the variable T maintains about the content information Y, and $\sum_{i=1}^{k} I(T;W_i)$ measures the information contained in variable T about the visual contexts $W_1, W_2, \cdots, W_k$. Once the joint distribution $p(X, Y)$ and visual contexts $W_1, W_2, \cdots, W_k$ are given, the terms $I(X;Y)$ and $\sum_{i=1}^{k} I(X;W_i)$ are constant and can be ignored. Thus, the objective function of CMIB can be written as follows:

$$\begin{aligned}\mathcal{L}_{max}\{p(t|x)\} &= \mathcal{I}^{G_{out}} - \beta^{-1} \cdot \mathcal{I}^{G_{in}} \\ &= I(T;Y) + \sum_{i=1}^{k} I(T;W_i) - \beta^{-1} I(X;T),\end{aligned} \tag{6}$$

where β strikes a balance between the information preservation and data compression.

The remaining task of the unsupervised object categorization is to maximize the value of the objective function (6). Obviously, the terms in the objective function (6) cannot be directly calculated, thus, we propose a novel sequential information-theoretic optimization to make them computable. The proposed optimization is essentially a sequential "draw-and-merge", which always performs better than agglomerative methods [30], especially when dealing with larger datasets, as shown in Section III-F. In this work, we concentrate on the "hard" clustering setting, where $p(t|x)$ is 0 or 1. Now, the goal of CMIB becomes to maximize the objective function (6).

C. Optimization of CMIB

To maximize the objective function (6) of CMIB, we propose a sequential information-theoretic optimization, which is essentially a "draw-and-merge" procedure. The draw-and-merge optimization is performed by the following steps:

1) Random initialization. The source image collection X is partitioned into M categories $T = \{t_1, \cdots, t_M\}$ stochastically. The mapping from X to T is denoted by $p(t|x)$, where $p(t|x) = 1$ means x belongs to the category t, and $p(t|x) = 0$ means x does not belong to t.

2) Draw. Each image $x \in X$ is drawn in order from its current category and is treated as a singleton category $\{x\}$.

3) Merge. To ensure that the total number of categories is M, the $\{x\}$ should be merged into category t^{new}. Let $\Delta\mathcal{L}(\{x\}, t)$ be the exact difference between the value of the objective function (6) before and after the merge. Since CMIB maximizes (6), $\{x\}$ should be merged into category $t^{new} = argmin_{t \in T} \Delta\mathcal{L}(\{x\}, t)$.

4) Convergence check. Repeat 2) and 3) until every x is not allocated to new category.

Now, the key issue is to compute the value change $\Delta\mathcal{L}(\{x\}, t)$ before and after x is merged into some category $t \in T$. Next, we give the probability computation when x is merged into a certain category $t \in T$ as the following definition.

Definition 3. *Suppose a certain $\{x\}$ is merged into some category t, and thus generates another new category t', the probability change caused by the merge step is defined as*

$$\begin{cases} p(t') = p(x) + p(t), \\ p(y|t') = \dfrac{p(x)}{p(t')} p(y|x) + \dfrac{p(t)}{p(t')} p(y|t), \end{cases} \tag{7}$$

where $p(t')$ is the prior probability after merging $\{x\}$ into category t, $p(y|t')$ is the joint probability distribution of category t' over the relevant information Y.

Let $\mathcal{L}^{bef}$ and $\mathcal{L}^{new}$ represent the value of Eq. (6) before and after the single category x is allocated to the category t. Then the value change $\Delta\mathcal{L}(\{x\}, t)$ can be calculated as follows:

$$\begin{aligned}&\Delta\mathcal{L}(\{x\}, t) = \mathcal{L}^{bef} - \mathcal{L}^{new} = [I(T^{bef};Y) - I(T^{new};Y)] + \\ &\sum_{i=1}^{k} [I(T^{bef};W_i) - I(T^{new};W_i)] - \\ &\beta^{-1}[I(T^{bef};X) - I(T^{new};X)] \\ &= \Delta I_{content} + \sum_{i=1}^{k} \Delta I_{context}^{i} - \beta^{-1} \Delta I_{compress},\end{aligned} \tag{8}$$

where T^{bef} and T^{new} are the categories before and after x is merged. $\Delta I_{content}$ is the value change of the CMIB objective function caused by content information Y. $\Delta I_{context}^{i}$ is the value change caused by the i-th visual context W_i. $\Delta I_{compress}^{i}$ is the value change caused by compressing X

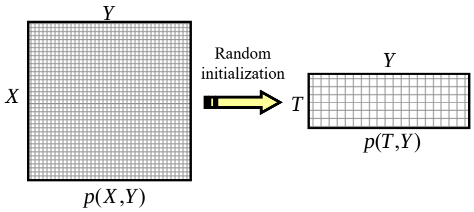


Fig. 4. The random initialization in the draw-and-merge optimization.

into its compressed representation T. From Eq. (8), we can see that the total value change $\Delta\mathcal{L}(\{x\},t)$ can be obtained from the calculation of $\Delta I_{content}$, $\Delta I^{i}_{context}$ and $\Delta I_{compress}$. First, according to Definition 2 and the definition of the mutual information, $\Delta I_{content}$ can be computed as follows:

$$\begin{aligned}
&\Delta I_{content} = I(T^{bef};Y) - I(T^{new};Y) \\
&= p(x)\sum_{y} p(y|x)log\frac{p(y|x)}{p(y)} + p(t)\sum_{y} p(y|t)log\frac{p(y|t)}{p(y)} \\
&- p(t')\sum_{y} p(y|t')log\frac{p(y|t')}{p(y)} \\
&= p(x)\sum_{y} p(y|x)log\frac{p(y|x)}{p(y)} + p(y)\sum_{y} p(y|t)log\frac{p(y|t)}{p(y)} \\
&- \sum_{y} p(x)p(y|x)log\frac{p(y|t')}{p(y)} - \sum_{y} p(t)p(y|t)log\frac{p(y|t')}{p(y)} \\
&= p(x)\sum_{y} p(y|x)log\frac{p(y|x)}{p(y|t')} + p(t)\sum_{y} p(y|t)log\frac{p(y|t)}{p(y|t')} \\
&= p(x)D_{KL}[p(y|x)||p(y|t')] + p(t)D_{KL}[p(y|t)||p(y|t')] \\
&= [p(x)+p(t)]JS_{\Pi}[p(y|x),p(y|t)].
\end{aligned} \tag{9}$$

where JS_{Π} is the *Jensen-Shannon (JS)* divergence [28] to compute the distance of two distributions, and $\Pi = \{\pi_1, \pi_2\} = \{\frac{p(x)}{p(x)+p(t)}, \frac{p(t)}{p(x)+p(t)}\}$. Since $p(t') \geq 0$ and $JS_{\Pi} \geq 0$, thus, $\Delta I_{content} \geq 0$. Similar analysis yields

$$\Delta I_{compress} = [p(x)+p(t)]JS_{\Pi}[p(x),p(x|t)]. \tag{10}$$

Next, we present how to initially compute the probabilistic terms in Eq. (9) and (10). As shown in Fig. 4, once given the joint distribution $p(X,Y)$, we can partition X into M categories $T = \{t_1, \cdots, t_M\}$ stochastically and obtain a initialized category partition $p(T,Y)$ by the random initialization in draw-and-merge optimization, where $p(t) = \sum_{x \in t} p(x)$. Now, all the probabilistic terms in these two equations can be computed from $p(X,Y)$ and $p(T,Y)$ as follows: the marginal probability $p(x) = \sum_{y \in Y} p(x,y)$, $p(t) = \sum_{x \in t} p(x)$; the conditional probability $p(y|x) = \frac{p(x,y)}{p(y)}$, $p(y|t) = \frac{p(y,t)}{p(t)}$, $p(x|t) = \frac{p(t|x)p(x)}{p(t)}$.

D. Relatedness Between Content and Visual Contexts

Now, we present the calculation of $\Delta I^{i}_{context}$. The sequential draw-and-merge procedure is an iterative procedure in essence. We use $T^{mid} = \{t^{mid}_1, t^{mid}_2, \cdots, t^{mid}_M\}$ to present the temporary partition in a certain iteration of CMIB conducted by the content variable, where M is the number of categories. Similarly, let W^l be the *l-th* clustering partition of multiple auxiliary visual contexts $W_1, W_2, \ldots, W_k$, which takes values from $W^l = \{w^l_1, w^l_2, \cdots, w^l_M\}$. To measure the relationship between the content information and the visual contexts, we construct the co-occurrence matrix between T^{mid} and W^l.

Algorithm 1 CMIB Algorithm

```
Input:
    Joint distribution $p(X,Y)$
    Multiple visual contexts $W_1, \cdots, W_k$
    Trade-off parameter $\beta$
    Cardinality value $M$
Random initialization: Stochastically divide the image collection $X$ into $M$ categories $T = \{t_1, t_2, \cdots, t_M\}$; Calculating the marginal probability $p(x) = \sum_{y \in Y} p(x,y)$, $p(t) = \sum_{x \in t} p(x)$; the conditional probability $p(y|x) = \frac{p(x,y)}{p(y)}$, $p(y|t) = \frac{p(y,t)}{p(t)}$, $p(x|t) = \frac{p(t|x)p(x)}{p(t)}$.
repeat
    for each image $x \in X$ do
        Draw: draw image $x$ from its original category.
        Calculate merge costs $\Delta\mathcal{L}(\{x\},t)$ by Eq. (8), which is computed from Eq. (9), Eq. (10) and Eq. (12).
        Merge: Allocate image $x$ into a new category $t^{new}$ that should satisfy $t^{new} = \arg\min_{t \in T} \Delta\mathcal{L}(\{x\},t)$.
    end for
until Every $x$ is not allocated into new category
Output: The final category $T$ of image collection $X$
```

As mentioned earlier, there are n images in the source data $X = \{x_1, x_2, \ldots, x_n\}$. Let n_i be the number of images allocated into category t^{mid}_i $(1 \leq i \leq M)$. Let n_j be the number of images allocated into category w^l_j $(1 \leq j \leq M)$. Let n_{ij} be the number of images allocated into category t^{mid}_i and w^l_j at the same time. Thus, the joint co-occurrence distribution of categories T^{mid} and visual context W^l can be computed as follows:

$$p(t^{mid}_i) = \frac{n_i}{n}, p(t^{mid}_j) = \frac{n_j}{n}, p(t^{mid}_i, w^l_j) = \frac{n_{ij}}{n}. \tag{11}$$

The mutual information of T^{mid} and W^l can be computed:

$$I(T;W^l) = \sum_{i=1}^{M}\sum_{j=1}^{M} p(t^{mid}_i, w^l_j) \log \frac{p(t^{mid}_i, w^l_j)}{p(t^{mid}_i)p(t^{mid}_j)}. \tag{12}$$

Thus, the total value change in Eq. (8) can be obtained. The pseudocode of CMIB is given in Algorithm 1. Next, we present how to deal with a fresh unseen data point that is available only after the optimization. We use x_{new} to indicate a fresh unseen data point. First, the fresh data point x_{new} is transformed into a co-occurrence vector by the bag-of-visual-word model. Then, x_{new} is randomly allocated into one cluster after the optimization of Algorithm 1 is finished. Now, the proposed draw-and-merge optimization is performed again until the objective function converges. Note that, since the

TABLE I
DETAILS OF THE SEVEN IMAGE DATASETS FOR OUR EXPERIMENTS.

Datasets	Category size	Data size	Codebook size
Soccer	7	280	1000
A-Yahoo	11	1100	1000
Dslr	31	489	1000
Webcam	31	795	1000
Amazon	31	2813	1000
CIFAR-10	10	15000	1000
NUS-22	22	10500	1000

(a) Soccer (b) NUS-22

Fig. 5. Example image categories from the Soccer and NUS-22 datasets.

visual contexts in this study is the group-partition information for an image dataset, there is no visual contexts for a singleton fresh data point. Thus, the merger cost is calculated just according to the Eq. (9) and Eq. (10) for the fresh data x_{new}.

E. Theoretical Analysis

1) Convergence: In this section, we prove the convergence of the objective function of the CMIB algorithm from upbound and monotone increasing. Next, we give Theorem 1.

Theorem 1. *The objective function of the CMIB algorithm can converge to a stable solution.*

Proof. First, we prove each draw-and-merge procedure do not decrease the value of the objective function. In CMIB, since the single category $\{x\}$ is allocated to the category t^{new} that $t^{new} = \arg\min_{t\in T} \Delta\mathcal{L}$, there must exist information loss once x is merged into the new categories, i.e., $\Delta\mathcal{L} \geq 0$. In other words, the "draw-and-merge" optimization will increase the function (6). Second, we prove the objective function (6) is upper bounded. The objective function (6) can be divided int two parts: $I(T;Y) - \beta^{-1}I(X;T)$ and $\sum_{i=1}^{k} I(T;W_i)$. The first part is the objective function of original IB, and its upbound was proven in [31]. For the second part, we obtain $I(T;W_i) \leq I(T;W)$ by assuming X has a true clustering W. Thus, the objective function (6) is upper bounded. Therefore, the objective function of CMIB can converge to a stable solution. □

2) Complexity: In this section, we analyze the time complexity of CMIB. In step 2, the source data X is partitioned into different categories with random initialization, so this step takes $O(1)$. In the main loop, the complexity of drawing data point x at step 5 is $O(1)$. The computation of the merge cost in step 6 takes $O(M|Y|)$, where M is the number of categories, and $|Y|$ is the dimension of the content feature. Therefore, the overall time complexity of CMIB is $O(M|X||Y|)$, where $|X|$ is the volume of data points. In the next experimental section, we will show the convergence of the objective function (6).

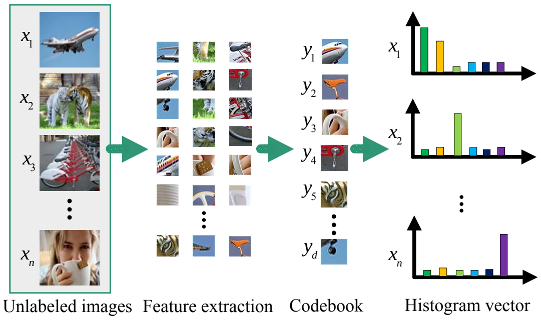


Fig. 6. Image representation with the bag-of-visual-words model.

IV. EXPERIMENTS

In this section, we conduct experiments to demonstrate the effectiveness of the proposed CMIB.

A. Image datasets

In our experiments, seven real-world image datasets including different object categories are employed for evaluation: Soccer [32], A-Yahoo [33], Dslr, Webcam and Amazon [34]. The detailed information is shown in Table I. Some example images in Soccer and NUS-22 are shown in Fig. 5.

Soccer [1] contains images collected from 7 soccer teams, containing 40 images per class, for a total of 280 images.

A-Yahoo [2] consists of 12 objects which were collected from the Yahoo image search. We select 11 categories in our experiments, with each category containing 100 images.

Dslr, Webcam and Amazon [3] consist of images captured with a digital SLR camera, recorded with a simple webcam and download from www.amazon.com, respectively. These three image datasets contain 31 object categories, which make them quite challenging to be partitioned due to the high diversity of the object categories.

CIFAR-10 [4] (Canada Institute For Advanced Research) consists of 15000 32x32 color images in 10 classes, with 1500 images per class.

NUS-22 [5] consists of online images and the associated tags, with a total number of 81 unique categories. We select 22 categories with a total number of 10500 images in our experiment.

B. Data representation

We exploit the popular bag-of-visual-words (BoVW) model, which has been widely used in various vision tasks, to represent the image collections. In the BoVW model, the following steps should be implemented:

[1] http://lear.inrialpes.fr/people/vandeweijer/data
[2] http://vision.cs.uiuc.edu/attributes/
[3] https://people.eecs.berkeley.edu/ jhoffman//domainadapt
[4] http://www.cs.toronto.edu/ kriz/cifar.html
[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

TABLE II
THE AC (%) COMPARISON OF CMIB WITH THE ORIGINAL IB AND FOUR OTHER TYPICAL CLUSTERING METHODS.

Datasets	IB			k-means	NCuts	pLSA	LDA	CMIB
	SIFT	Color Attention	TPLBP					
Soccer	35.86 ± 1.9	**52.07 ± 4.2**	23.68 ± 1.3	43.14 ± 2.2	49.46 ± 0.2	47.04 ± 3.9	49.04 ± 4.0	**56.04 ± 3.2**
A-Yahoo	26.88 ± 1.1	19.11 ± 0.8	**33.19 ± 1.2**	23.43 ± 1.3	21.43 ± 0.2	31.39 ± 1.4	32.93 ± 0.9	**36.26 ± 1.6**
Dslr	**42.81 ± 1.5**	34.45 ± 1.3	39.72 ± 2.0	32.63 ± 2.0	31.49 ± 0.7	30.82 ± 1.5	33.01 ± 1.2	**49.32 ± 0.8**
Webcam	**38.84 ± 1.3**	28.62 ± 1.2	36.28 ± 1.0	20.84 ± 1.2	31.56 ± 0.6	29.60 ± 1.7	28.73 ± 1.1	**42.18 ± 2.6**
Amazon	24.48 ± 0.8	12.56 ± 0.6	**26.51 ± 1.2**	13.33 ± 0.5	15.08 ± 0.2	18.85 ± 1.4	15.86 ± 1.3	**29.17 ± 1.3**
CIFAR-10	18.57 ± 0.9	**20.13 ± 0.8**	19.79 ± 1.1	13.88 ± 0.4	17.36 ± 0.5	15.86 ± 0.8	15.98 ± 0.2	**25.44 ± 1.5**
NUS-22	13.12 ± 1.0	12.42 ± 0.7	**17.57 ± 1.2**	15.72 ± 0.6	18.55 ± 0.4	16.90 ± 0.3	14.76 ± 0.5	**21.84 ± 1.5**
Average	28.65	25.62	28.11	23.28	26.42	27.21	27.19	**37.18**

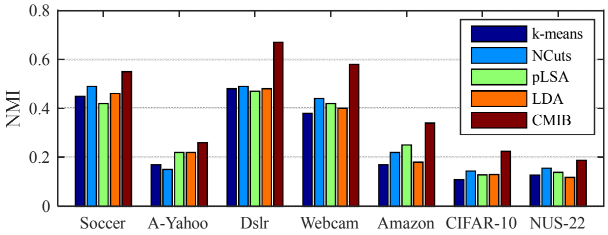


Fig. 7. The NMI comparison of CMIB with typical clustering methods.

1) Feature detection. We first detect the interest points for each image. In particular, Harries corner detection [35] is adopted to detect the local interest points, while a dense sampling method [36] is utilized to capture the dense interest points for the global image features. In this study, 128 dimensional SIFT [35], 36 dimensional Color Attention [37] and 256 dimensional TPLBP [38] descriptors are utilized to represent the content feature, global color and texture features, respectively.

2) Codebook generation. A visual codebook is built using the k-means algorithm and Euclidean distance as the clustering metric for each type of descriptor. We set the number of clusters as 1000, i.e, there are 1000 visual words in each codebook.

3) Mapping descriptors into the codebook. All the descriptors are mapped into their corresponding visual word index.

4) Counting occurrence. The occurrence number of each visual word is counted for each image, thus each image can be represented by a histogram vector including occurrence number of visual words.

Fig. 6 presents an illustration of the image representation with the BoVW model. Given an unlabeled image collection $X = \{x_1, \cdots, x_n\}$, we can obtain the visual codebook $Y = \{y_1, \cdots, y_d\}$ by the BoVW model, where n and d are the number of images and visual words in the codebook, respectively. Based on the BoVW model, each image can be transformed into a histogram vector, thus, we obtain the conditional distribution of the visual words as $p(y|x) = \frac{n(y|x)}{\sum_{y' \in Y} n(y'|x)}$, where $n(y|x)$ is the number of occurrences of the word y in the image x. To avoid an undesirable bias due to different numbers of features in images, we set a uniform prior distribution $p(x) = \frac{1}{n}$. Thus, the joint distribution between the source image variable X and the visual word variable Y (we also call it a relevant variable in this study) can be computed by $p(x, y) = p(y|x)p(x)$. In this study, we designate the SIFT as the content information of the object, while the global color and texture features are utilized to generate the visual contexts.

C. Comparison methods

We adopt five types of comparison methods: 1) Information bottleneck. 2) Traditional clustering methods: k-means, normalized cuts (NCuts) [39], probabilistic latent semantic analysis (pLSA) [40], and latent Dirichlet allocation (L-DA) [41]. 3) Ensemble clustering methods: cluster-based similarity partitioning algorithm (CSPA), hyper-graph partitioning algorithm (HGPA), meta-clustering algorithm (MCLA) [23] and locally weighted evidence accumulation (LWGP) [24]. 4) Multi-view clustering methods: coregularized multi-view spectral clustering (CRSC) [20], cotraining multi-view spectral clustering (CTSC) [19], robust multi-view spectral clustering (RMSC) [21] and multivariate information bottleneck (MIB) [27]. 5) Image clustering methods: local discriminant models and global integration (LDMGI) [42], clustering-by-composition (CC) [43] and ensemble projection (EP) [44].

To ensure the fairness of comparisons, all the input datasets are represented by the widely used BoVW model. The total number of clusters is provided for all the clustering algorithms. To fairly compare performance of IB, MIB and CMIB, we fix the trade-off parameter β as 40. For NCuts, LDMGI and CC, the number of nearest neighbors is fixed as 5 according to [42], [43], and the best regularization parameter in LDMGI is searched in $\{10^{-8}, 10^{-6}, \cdots, 10^{6}, 10^{8}\}$. For LDA, the Dirichlet parameter $\alpha = M/50$, where M is the number clusters of each dataset. For CTSC, CRSC and RMSC, Gaussian kernels are used to build the similarity matrix for each single view, and the regularization parameter λ is varied from 0.01 to 0.05 and the best result is reported. For LWGP, the cluster uncertainty parameter $\theta = 0.4$ for all datasets. For EP, we set the number of weak training sets as 1000, where the training sets are created by performing a random walk sampling [44] on the unlabeled images, and the number of training images for each class is set as 9. In this study, we use the ground-truth label to evaluate the final clustering quality of all baselines, and all clustering stages do not involve looking at the ground-truth label. Besides, the number of clusters need to be specified up front for all the baselines used in this study including CMIB.

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/TII.2019.2939278, IEEE Transactions on Industrial Informatics

TABLE III
THE AC (%) COMPARISON OF CMIB WITH STATE-OF-THE-ART ENSEMBLE AND MULTI-VIEW CLUSTERING METHODS.

Datasets	Ensemble clustering				Multi-view clustering				CMIB
	CSPA	HGPA	MCLA	LWGP	CTSC	CRSC	RMSC	MIB	
Soccer	53.93 ± 0.1	39.89 ± 6.0	47.11 ± 0.3	50.69 ± 2.3	38.39 ± 3.7	31.46 ± 1.9	27.04 ± 1.8	51.12 ± 3.0	**56.04 ± 3.2**
A-Yahoo	32.79 ± 0.3	20.75 ± 1.4	32.17 ± 0.6	32.46 ± 1.2	31.43 ± 1.1	29.00 ± 0.6	25.20 ± 0.4	32.17 ± 1.4	**36.26 ± 1.6**
Dslr	45.58 ± 0.5	39.38 ± 1.5	42.25 ± 0.4	41.74 ± 2.1	41.29 ± 1.5	36.08 ± 1.3	35.92 ± 1.5	44.26 ± 1.0	**49.32 ± 0.8**
Webcam	40.75 ± 0.1	33.19 ± 2.1	37.42 ± 0.7	39.85 ± 2.2	38.25 ± 2.0	36.31 ± 0.8	28.74 ± 1.1	39.28 ± 2.3	**42.18 ± 2.6**
Amazon	25.19 ± 0.6	14.67 ± 0.7	22.85 ± 0.4	23.37 ± 1.5	26.08 ± 1.0	21.32 ± 0.6	15.43 ± 0.5	20.62 ± 1.1	**29.17 ± 1.3**
CIFAR-10	23.68 ± 0.3	16.16 ± 1.2	18.66 ± 0.5	18.78 ± 2.0	21.02 ± 1.3	18.66 ± 0.8	20.80 ± 1.0	21.15 ± 1.2	**25.44 ± 1.5**
NUS-22	20.72 ± 0.4	18.55 ± 1.6	17.90 ± 0.7	16.76 ± 1.6	18.63 ± 1.4	19.28 ± 0.7	17.36 ± 0.9	19.38 ± 1.6	**21.84 ± 1.5**
Average	34.67	26.08	31.45	31.95	30.73	27.44	24.36	32.57	**37.18**

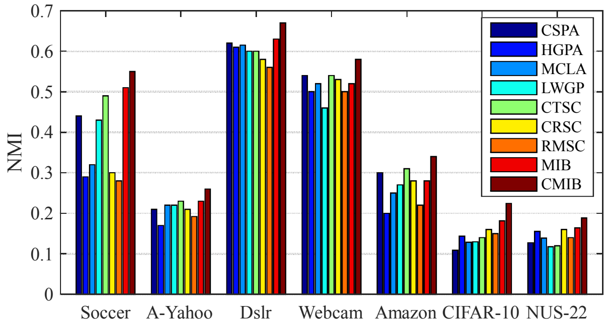


Fig. 8. The NMI comparison of CMIB with state-of-the-art ensemble and multi-view clustering methods.

D. Evaluation metrics

In this paper, we employ two metrics, normalized mutual information (NMI) and clustering accuracy (AC) [19], to evaluate the performance of the different methods.

Given an image x_i, let l_i and t_i be the ground-truth label and the obtained cluster label respectively, so AC is:

$$AC = \frac{\sum_{i=1}^{n} \delta(l_i, \mathrm{map}(t_i))}{n}, \tag{13}$$

where n is the total number of images and $\delta(l_i, \mathrm{map}(t_i))$ is the delta function that equals 1 if $x = y$, otherwise the delta function equals 0, and $map(t_i)$ is the optimal mapping function that permutes clustering labels to match the ground-truth labels. The optimal mapping is obtained by the Kuhn-Munkres algorithm [45].

Unlike AC, NMI is an information theoretic-based metric [23], that estimates the quality of the clustering results by measuring the degree of agreement between the learned clusters and the ground-truth class. NMI can be estimated by

$$NMI = \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}, \tag{14}$$

where n_i indicates the number of images in cluster C_i ($1 \leq i \leq c$), $\hat{n}_j$ is the number of images in ground-truth class L_j ($1 \leq j \leq c$), and $n_{i,j}$ indicates the number of images that are in the intersection between cluster C_i and class L_j. The larger the value of NMI the better the clustering results will be.

TABLE IV
THE AC (%) COMPARISON OF CMIB WITH STATE-OF-THE-ART IMAGE CLUSTERING METHODS.

Datasets	LDMGI	CC	EP	CMIB
Soccer	43.21 ± 0.5	47.75 ± 2.2	48.21 ± 0.4	**56.04 ± 3.2**
A-Yahoo	28.11 ± 1.1	32.27 ± 0.6	30.87 ± 1.5	**36.26 ± 1.6**
Dslr	37.51 ± 1.6	47.27 ± 1.7	36.22 ± 1.6	**49.32 ± 0.8**
Webcam	33.53 ± 1.0	40.97 ± 0.8	24.47 ± 0.6	**42.18 ± 2.6**
Amazon	19.56 ± 0.7	26.04 ± 1.0	20.81 ± 0.5	**29.27 ± 1.3**
CIFAR-10	18.05 ± 3.4	19.57 ± 1.2	20.79 ± 0.1	**25.44 ± 1.5**
NUS-22	17.09 ± 0.1	18.11 ± 0.9	16.32 ± 0.8	**21.84 ± 1.5**
Average	28.15	33.14	28.24	**37.18**

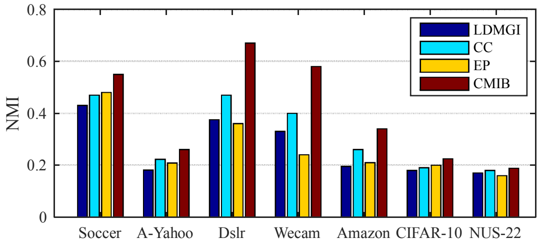


Fig. 9. The NMI comparison of CMIB with state-of-the-art image clustering methods.

E. Comparative Analysis

In this part, extensive experiments are conducted to demonstrate the effectiveness of CMIB compared with seven types of comparison methods.

1) Baselines without visual contexts: We conduct experiments to verify the performance of CMIB compared with the original IB method. From Table II, we have the following observations. First, the clustering results (AC) of the IB algorithm on the three cues are different. This demonstrates that a single type of feature is not sufficiently discriminative and stable for different datasets. Second, by incorporating multiple visual contexts, the proposed CMIB method clearly performs better than the IB algorithm.

Further experiments are conducted to compare the CMIB with four other traditional clustering approaches: k-means, N-Cuts [39], pLSA [40] and LDA [41]. From Table II and Fig. 7, we observe that the CMIB method significantly outperforms all other traditional clustering methods, which is mainly caused by the visual contexts used in our CMIB method.

2) Comparison with ensemble and multi-view clustering methods: Essentially, contextual information is a complement

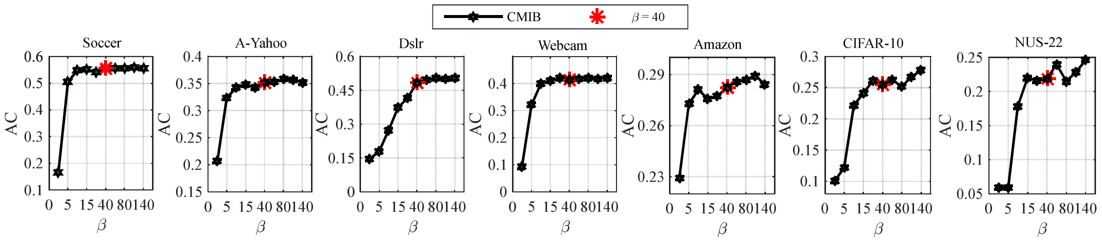


Fig. 10. The AC values of CMIB on the seven datasets with different trade-off parameters.

to the content information of an object. In this regard, it is pertinent to discuss multi-view and ensemble clustering methods. Both aim to improve the performance of the unsupervised object categorization model by considering the complementary effect of multiple related components.

In this section, we first conduct comparative experiments with ensemble clustering methods (CSPA, HGPA, MCLA [23] and LWGP [24]) to demonstrate the effectiveness of CMIB. We utilize the original IB to construct 30 base clusterings for the above ensemble clustering methods, in which each feature generates 10 clusterings. From Table III and Fig. 8, the CMIB outperforms all four ensemble clustering methods on the seven datasets. This is mainly because ensemble clustering methods usually limit the final results to the quality of the existing base clusterings. The proposed CMIB method can deal with the original feature (content feature) and visual contexts (basic clusterings) simultaneously and relieve the overreliance of ensemble clustering methods on auxiliary clusterings.

To further demonstrate the performance of CMIB, we compare it with other four multi-view clustering methods: CTSC [19], CRSC [20], RMSC [21] and MIB [46]. In this experiment, we treat each feature (SIFT, Color Attention and TPLBP) as one input view of the multi-view clustering methods. From Table III and Fig. 8, we can see that the performances of CMIB are also significantly better than the multi-view clustering methods. The CMIB algorithm also outperforms the original MIB, mainly because the original MIB can only address information sources with completely homogenous data distributions, while the multiple visual contexts in the proposed CMIB naturally have the ability to address heterogenous features.

3) Comparison with state-of-the-art image clustering methods: For comparison with promising unsupervised object categorization methods, we adopt local discriminant models and global integration (LDMGI) [42], clustering-by-composition (CC) [43] and ensemble projection (EP) [44] as baselines. As shown in Table IV, the average AC values of the CMIB algorithm on the seven datasets obtain 17.47%, 13.98%, and 17.41% improvements, compared with the other three baselines. We have the same observations in terms of NMI in Fig. 9.

F. Explanation of impact factors

1) Parameter analysis: In CMIB, β strikes a balance between information preservation and data compression. Thus, we conduct an experiment to investigate the impact of β on the performance of CMIB. Specifically, we vary the values of β from the space $\{1, 5, 10, 15, 20, 40, 60, 80, 100, 140\}$. From Fig. 10, we obtain the following observations: First, when $\beta \to 0$, CMIB performs poorly since it only addresses the compression from source images X to its compressed representation T, i.e., the object category in this study. When increasing the value of β, CMIB performs much better because it strikes a balance between the data compression and information preservation. We set β to 40 on all the datasets in this study.

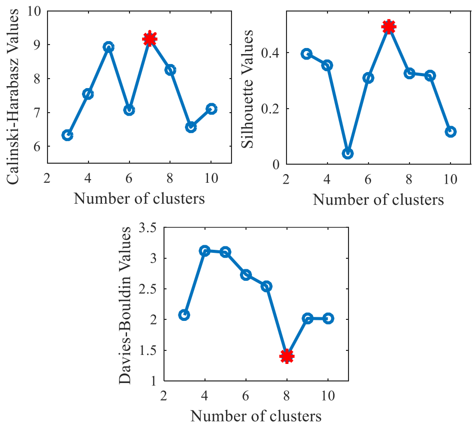


Fig. 11. The relationship between internal CVIs value of the CMIB and the number of categories. Note that, the larger CH and SI value and the smaller DB value are preferred. The red points indicate the peak values of CH SI and DB.

2) The influence of different numbers of categories: In real world clustering applications, the number of categories M is often unknown. Although the external cluster validity indices (CVIs) can accurately evaluate the quality of a clustering by measuring the similarity or dissimilarity between the ground-truth and candidate categories, it is impossible to generate the ground-truth with different numbers of clusters for one dataset. Thus, we resort to internal cluster validity indices (CVIs) [47] to measure the quality of clustering with different numbers of clusters, which evaluate the quality of clustering results based only on the data themselves. As evaluation metrics without ground-truth label, we adopt Calinski-Harabasz values (CH), Silhouette values (SI), Davies-Bouldin values (DB), as they are widely used in the literature.

Fig. 11 presents the CH, SI and DB values of CMIB on the Soccer dataset by varying the number of categories from 3 to 10. As shown in this figure, we obtain the following

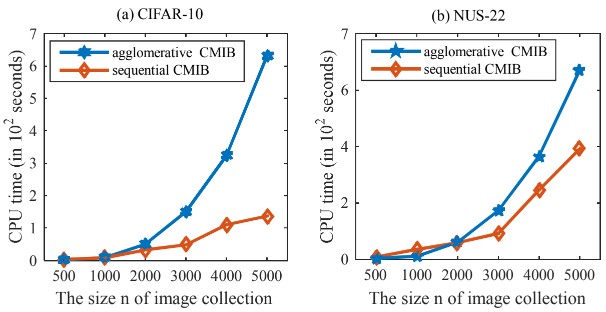


Fig. 12. Time complexities with respect to different sizes of CIFAR-10 and NUS-22 datasets.

TABLE V
THE AC (%) VALUE OF CMIB ON DIFFERENT MULTITUDES OF FEATURES. THE THREE COLUMNS ARE THE RESULTS OF CMIB WITH THE VISUAL CONTEXTS BY ONE AND TWO TYPES OF GLOBAL FEATURES. THE UNDERLINED RESULTS SHOW THE SECOND BEST RESULTS.

Datasets	CMIB		
	SIFT+Color	SIFT+TPLBP	SIFT+Color+TPLBP
Soccer	53.46 ± 2.8	38.23 ± 2.2	**56.04** ± **3.2**
A-Yahoo	29.17 ± 1.4	35.64 ± 1.8	**36.26** ± **1.6**
Dslr	45.63 ± 1.1	47.38 ± 0.9	**49.32** ± **0.8**
Webcam	40.12 ± 2.1	40.16 ± 1.8	**42.18** ± **2.6**
Amazon	26.43 ± 0.9	28.42 ± 1.4	**29.27** ± **1.3**
CIFAR-10	22.56 ± 1.2	20.08 ± 1.8	**25.44** ± **1.5**
NUS-22	15.23 ± 1.5	20.49 ± 1.3	**21.84** ± **1.5**
Average	33.23	32.91	**37.18**

observations. First, the values of the three CVIs fluctuates to a certain degree with varying the numbers of categories. Second, the corresponding cluster numbers are 7, 7 and 8 when the CH, SI and DB obtain the optimal value on the Soccer dataset. They are close to the number of the known genuine class of the Soccer dataset, which is 7. Thus, the internal CVIs can provide some guidance for the automatic determination of the numbers of categories.

3) Time complexity analysis: In this study, we propose a novel sequential information-theoretic solution to optimize the objective function of the CMIB algorithm. To compare the running time of the sequential solution with the agglomerative solution, we design an agglomerative optimization based on [30] to optimize CMIB and call it "agglomerative CMIB". To distinguish, we name the sequential solution "sequential CMIB". As shown in Fig. 12, the running times of the sequential and agglomerative CMIB are comparable when the data size is small. With an increase in data size, the time increment of the agglomerative CMIB is larger than that of the sequential one. As shown in Section III-E, the time complexity of the sequential CMIB is $O(M|X||Y|)$, where M, $|X|$ and $|Y|$ are the category number, data size and feature dimensionality. In clustering scenario, the category number is always far smaller than the number of datasets, i.e., $M << |X|$, thus, we can obtain significant run time improvement.

4) Multitude of feature representations: The proposed CMIB adopts a content feature (SIFT in this study) to characterize the content information of the target object, while automatically generating a set of visual contexts by multiple global features. Thus, we conduct experiments to show the impact of different multitudes of feature representations on the performance of CMIB. As shown in Table V, we obtain the following observations. First, the AC values of CMIB fluctuate to a certain degree when considering individual global feature. This is mainly caused by the distinguishing abilities of different features. For instance, the TPLBP features cannot depict the Soccer data very well (refer to Table II), therefore, the results of CMIB are not satisfactory results by considering SIFT + TPLBP on Soccer dataset. Second, the clustering performance of CMIB is superior and stable when considering two visual contexts. This also verifies the robustness of CMIB on multitude of possible feature representations. We believe the performance of CMIB can be further improved by more promising features, such as CNN feature.

5) The performance on classification metrics: In essence, both clustering and classification approaches aim to distinguish a set of data instances into categories. Specifically, the clustering approaches divide *unlabeled* data into a set of disjoint subsets with high intra-cluster similarity and low inter-cluster similarity, while the classification approaches first train a classifier using *labeled* data, then a new-coming unlabeled data instance is labeled by the classifier. Thus, some typical classification metrics can be adopted to evaluate the performance of clustering method once the ground-truth label is given, which is usually used in the literature of external cluster validity indices [47].

In this section, we adopt F_1-$Measure$ [48] to further evaluate the performance of CMIB. F_1-$Measure$ depicts the overall performance of the clustering or classification results, which refers to the harmonic average of $precision$ (i.e., the ratio of true positives to all instances predicted as positive) and $recall$ (i.e., the ratio of true positives to all instances that are actually positive). F_1-$Measure$ is defined as follows:

$$\begin{aligned} precision &= \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN}, \\ F_1 - Measure &= \frac{2 \cdot precision \cdot recall}{precision + recall}, \end{aligned} \tag{15}$$

where TP, FP and FN are the number of true positives, false positives and false negatives, respectively. The larger the value of F_1-$Measure$ the better the clustering results will be.

Fig. 13 presents the F_1-$Measure$ comparison of CMIB with all the baselines on the seven datasets. From Fig. 13, we can observe that the proposed CMIB algorithm performs much better than other baselines in terms of the F_1-$Measure$ metrics. The comparison with typical classification metrics further demonstrates the effectiveness of the proposed CMIB algorithm, which also would open up a much wider audience for this study.

6) Convergence analysis: Fig. 14 shows the number of iterations of the CMIB algorithm on the seven datasets. It can be observed that function (6) increases monotonically and reaches convergence rapidly in a limited number of iterations.

V. CONCLUSIONS

We propose a novel contextual multivariate information bottleneck (CMIB) approach, which aims to discover the

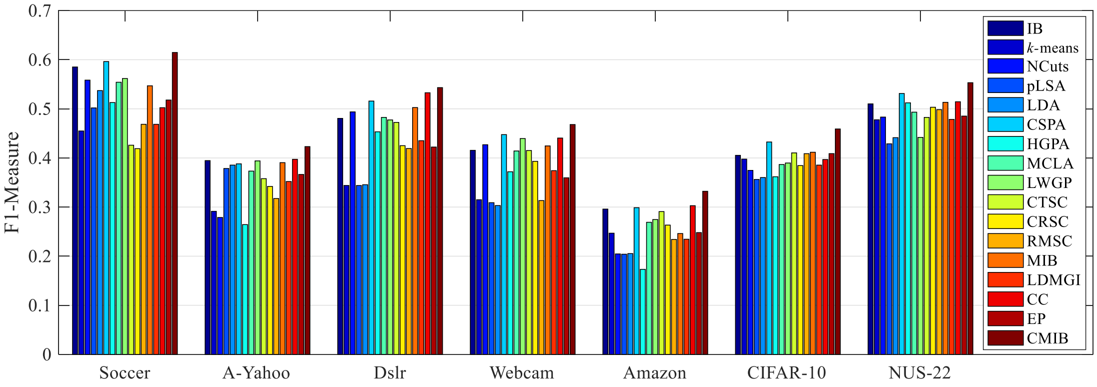


Fig. 13. The F_1-$Measure$ comparison of CMIB with all the baselines on the seven datasets used in this study.

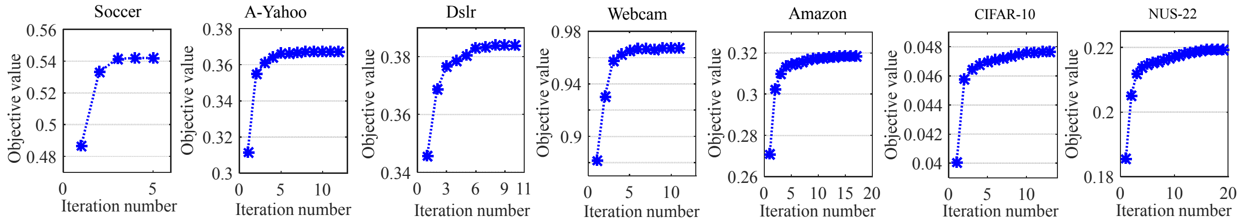


Fig. 14. The iterations of CMIB on the seven datasets.

object category in unlabeled images by simultaneously considering content information and their visual contexts. Rather than using the manual contexts, we focus on automatically constructing visual contexts from the unlabeled image data. CMIB treats object category discovery as a data compression procedure, in which both content and contextual information can be preserved maximally. Specifically, CMIB utilizes two Bayesian networks to characterize the relationship between data compression and information preservation. We present extensive experiments showing that the performance of our CMIB method is superior to other existing state-of-the-art baselines. In future research, we will investigate more meaningful visual contexts and test the proposed method on more realistic applications, for example, unsupervised domain adaptation via multi-task clustering.

References

[1] Y. Park and I. S. Kweon, "Ambiguous surface defect image classification of amoled displays in smartphones," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 597–607, 2016.

[2] J. Yang, B. Jiang, B. Li, K. Tian, and Z. Lv, "A fast image retrieval method designed for network big data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2350–2359, 2017.

[3] T.-T. Do and N.-M. Cheung, "Embedding based on function approximation for large scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 626–638, 2018.

[4] L. Zhao, Z. He, W. Cao, and D. Zhao, "Real-time moving object segmentation and classification from hevc compressed surveillance video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1346–1357, 2017.

[5] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019.

[6] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *International Conference on Computer Vision*, 2005, pp. 1284–1291.

[7] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[8] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 129–136.

[9] X. Song, S. Jiang, and L. Herranz, "Joint multi-feature spatial context for scene recognition on the semantic manifold," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1312–1320.

[10] L. Yang, K. Tang, J. Yang, and L. J. Li, "Dense captioning with joint inference and visual context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1978–1987.

[11] J. Yuan and Y. Wu, "Context-aware clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[12] H. Wang, J. Yuan, and Y. Tan, "Combining feature context and spatial context for image pattern discovery," in *International Conference on Data Mining*, 2011, pp. 764–773.

[13] V. Nguyen, D. Phung, X. Nguyen, and H. H. Bui, "Bayesian nonparametric multilevel clustering with group-level contexts," in *International Conference on Machine Learning*, 2014, pp. 288–296.

[14] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1352–1366, 2018.

[15] S. Jones and L. Shao, "Unsupervised spectral dual assignment clustering of human actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 604–611.

[16] F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Computer Vision and Pattern Recognition*, 2011, pp. 1977–1984.

[17] J. Wang, Z. Chen, and Y. Wu, "Action recognition with multiscale spatio-temporal contexts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3185–3192.

[18] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[19] A. Kumar and H. Daume, "A co-training approach for multi-view spectral clustering," in *International Conference on Machine Learning*, 2011, pp. 393–400.

[20] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral

clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[21] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2149–2155.

[22] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1774–1782, 2019.

[23] A. Strehl and J. Ghosh, "Cluster ensembles-a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

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

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

[26] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, "Multivariate information bottleneck," in *Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 152–161.

[27] N. Slonim, N. Friedman, and N. Tishby, "Multivariate information bottleneck," *Neural Computation*, vol. 18, no. 8, pp. 1739–1789, 2006.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1991.

[29] J. Philbin, O. Chum, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[30] N. Slonim, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*, 1999, pp. 617–623.

[31] ——, "The information bottleneck: Theory and applications," *Ph.D Dissertation, Hebrew University*, 2002.

[32] J. V. D. Weijer and C. Schmid, "Coloring local feature extraction," in *European Conference on Computer Vision*, 2006, pp. 334–348.

[33] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[34] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010, pp. 213–226.

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

[36] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision*, 2006, pp. 490–503.

[37] F. S. Khan, J. D. Weijer, and M. Vanrell, "Top-down color attention for object recognition," in *International Conference on Computer Vision*, 2009, pp. 979–986.

[38] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *European Conference on Computer Vision*, 2008.

[39] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[40] T. Hofmann, "Probabilistic latent semantic analysis," in *Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[41] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[42] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2761–2773, 2010.

[43] A. Faktor and M. Irani, "Clustering by composition - unsupervised discovery of image categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1092–1106, 2014.

[44] D. Dai and L. J. V. Gool, "Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering," *CoRR*, vol. abs/1602.00955, 2016.

[45] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Dover Publications, 1998.

[46] X. Yan, Y. Ye, and Z. Lou, "Unsupervised video categorization based on multivariate information bottleneck method," *Knowledge-Based Systems*, vol. 84, pp. 34–45, 2015.

[47] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 982–994, 2013.

[48] C. Goutte and É. Gaussier, "A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation," in *European Conference on Information Retrieval*, 2005, pp. 345–359.

Xiaoqiang Yan is a lecturer with School of Information Engineering at Zhengzhou University. He received his PhD degree in Software Engineering at Zhengzhou University. He received his B.S. degree and M.S. degree in School of Information Engineering from Zhengzhou University. He worked one year as a visiting scholar in University of Portsmouth, UK. His main research interests include computer vision, pattern recognition and data mining.

Yangdong Ye is a Professor with School of Information Engineering at Zhengzhou University. He received his Ph.D. degree in China Academy of Railway Sciences. He has wide research interests, mainly including machine learning, pattern recognition, knowledge engineering and intelligent system. He has published several papers in peer-reviewed prestigious journals and conference proceedings, such as IEEE Transactions on Multimedia, Neural Networks, IEEE CVPR, IJCAI and ACM Multimedia. More details about his research and background can be found at http://www5.zzu.edu.cn/mlis/

Xueying Qiu received her B.S. degree in School of Computer Science and Technology, Northeast Forestry University. She received her M.S. degree in School of Information Engineering from Zhengzhou University. Her main research interests include computer vision, machine learning and artificial intelligence.

Milos Manic (SM06-M04-StM96) received the Dipl.Ing. and M.S. degrees in electrical engineering and computer science from the University of Niš, Niš, Serbia in 1991 and 1997 respectively, and the Ph.D. degree in computer science from the University of Idaho in 2003.

Dr. Manic is a Professor with Computer Science Department and Director of VCU Cybersecurity Center at Virginia Commonwealth University. Dr. Manic has given over 40 invited talks around the world, authored over 200 refereed articles in international journals, books, and conferences, holds several U.S. patents and has won 2018 R&D 100 Award for Autonomic Intelligent Cyber Sensor (AICS). He is an officer of IEEE Industrial Electronics Society, founding chair of IEEE IES Technical Committee on Resilience and Security in Industry, and general chair of IEEE IECON 2018, IEEE HSI 2019.

Hui Yu is Professor with the University of Portsmouth, UK. Prof. Yu received PhD from Brunel University London. He used to work at the University of Glasgow before moving to the University of Portsmouth. His research interests include methods and practical development in vision, machine learning and AI with applications to human-machine interaction, Virtual and Augmented reality, robotics and geometric processing of facial expression. He serves as an Associate Editor of IEEE Transactions on Human-Machine Systems and Neurocomputing journal.