Combining Feature Context and Spatial Context for Image Pattern Discovery

Hongxing Wang, Junsong Yuan, Yap-Peng Tan
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
hwang8@e.ntu.edu.sg, [jsyuan, eyptan]@ntu.edu.sg

Abstract—Once an image is decomposed into a number of visual primitives, e.g., local interest points or salient image regions, it is of great interest to discover meaningful visual patterns from them. Conventional clustering (e.g., k-means) of visual primitives, however, usually ignores the spatial dependency among them, thus cannot discover the high-level visual patterns of complex spatial structure. To overcome this problem, we propose to consider both spatial and feature contexts among visual primitives for pattern discovery. By discovering both spatial co-occurrence patterns among visual primitives and feature co-occurrence patterns among different types of features, our method can better handle the ambiguities of visual primitives, by leveraging these co-occurrences. We formulate the problem as a regularized k-means clustering, and propose an iterative bottom-up/top-down self-learning procedure to gradually refine the result until it converges. The experiments of image texton discovery and image region clustering convince that combining spatial and feature contexts can significantly improve the pattern discovery results.

Keywords—clustering, image pattern discovery, feature context, spatial context.

I. INTRODUCTION

Images can be decomposed into visual primitives, e.g., interest point features or salient regions. As each visual primitive is usually described by a feature vector, it is of great interests to cluster these visual primitives into prototypes, e.g., visual words. Then by representing an image as a visual document, conventional text analysis methods can be directly applied. Despite many previous successes, it is a non-trivial problem to cluster visual primitives into an optimal visual vocabulary. For example, a simple k-means clustering of visual primitives can lead to synonymous visual words that over-represent visual primitives, as well as polysemous visual words that bring large uncertainties and ambiguities in the representation.

As visual primitives are not independent of each other, to better address the visual polysemous and synonymous phenomena, the ambiguities among visual primitives can be partially resolved through analyzing their spatial contexts [1], [2], i.e. other primitives in the spatial neighborhood. For example, two visual primitives, although exhibit dissimilar visual features, may belong to the same pattern if they have the same spatial contexts. On the other hand, two visual primitives, even though similar to each other, may not belong to the same pattern if their spatial contexts are completely different.

Despite previous work of considering spatial contexts for visual primitive clustering, usually a visual primitive is characterized by a single type of feature only, e.g. the SIFT descriptor [3]. However, a visual primitive can be naturally represented by different types of features, such as color, texture, and shape, it is worth studying how to combine these features together for a better clustering of them, e.g., through multi-view clustering. Moreover, in terms of visual pattern discovery, besides the spatial dependency among visual primitives, a visual pattern would exhibit certain feature dependency as well. For example, for a visual pattern of a red circle, it has a co-occurrence of the color (red) and shape (circle) features. However, it remains an open problem to combine both spatial and feature contexts simultaneously to improve the performance of visual pattern discovery.

In order to consider the spatial and feature contexts of visual primitives, we propose to discover the spatial co-occurrence patterns of visual primitives, as well as the feature co-occurrence patterns of different types of features. Such co-occurrence patterns, once discovered, can be utilized to resolve the ambiguities and uncertainties among visual primitives. A novel regularized k-means clustering is proposed to take these co-occurrence patterns into consideration. When clustering the visual primitives, spatial co-occurrence pattern and feature co-occurrence pattern can boost each other for a better clustering of visual primitives. To extract the high-level visual patterns from low-level features, an iterative bottom-up/top-down procedure is proposed to gradually refine the clustering results until convergence.

We test our proposed method in both image texton discovery and unsupervised image region categorization [4]. The results show that by combining both spatial and feature contexts, our method can effectively discover the visual patterns of complex spatial structures. The proposed self-learning method can converge in a few number of iterations.

II. RELATED WORK

The proposed method can be used for multi-view clustering, although it is not the major focus of this study. Comparing with single-view clustering in a feature space, multi-view clustering focuses on learning a more accurate partition of instances under multiple feature spaces. Multi-view data are universal in many practical applications, and more and more attention has been given to this topic. Some
related work can be found in [5], [6], [7], [8], [9], [10], [11], [12], [13].

In this work, our concern is visual pattern discovery [1], [2]. We first perform a multi-view clustering. Afterwards with the help of spatial contexts [1], [2], the clustering results are expected to improve the performance of visual pattern discovery. Actually spatial contextual information is often utilized to refine image feature description and representation [14], [15]. In this study, it is used to combine multi-view features for clustering performance improvement and visual pattern disambiguation.

Our work is an extension of [1], which separately shows how spatial relations among visual primitives can assist individual image texton discovery and how multiple representations of visual primitives to improve clustering. Our method is performed jointly using multi-view modalities and spatial contextual information among visual primitives. Therefore it can better handle the ambiguities of visual primitives.

III. PROBLEM STATEMENT

Given a dataset \( \mathcal{D}_v = \{v_i\}_{i=1}^N \) of \( N \) visual primitives, e.g., local image patches or regions, each \( v_i \) has \( c \) types (views) of feature representations: \( \{f_i^{(j)}\}_{j=1}^c \), where \( f_i^{(j)} \in \mathbb{R}^d \).

Using one type of feature representations, traditional \( k \)-means clustering groups the visual primitives into several clusters based on a similarity measure (e.g., Euclidean distance) such that the ones with similar features are in the same cluster and those with dissimilar features are in different clusters. Suppose we choose the first type of features \( f_i^{(1)} \in \mathbb{R}^{i_1} \) \( (i = 1, 2, \ldots, N) \). In order to repository \( \mathcal{D}_v \) into \( M_1 \) clusters, \( k \)-means clustering requires to minimize the following mean square distortion:

\[
J^{(1)} = \sum_{m=1}^{M_1} \sum_{n=1}^{N} r_{mn}^{(1)} ||u_{mi}^{(1)} - f_{ni}^{(1)} ||^2 = tr(R_1^T D_1),
\]

where \( M_1 \) is the number of clusters, \( u_{mi}^{(1)} \in \mathbb{R}^{i_1} \) is the \( m \)th prototype of visual word; \( || \cdot ||^2 \) denotes the Euclidean square distance; \( tr(\cdot) \) denotes the matrix trace; \( D_1 \) is the \( M_1 \times N \) distortion matrix, the entry \( d_{mn}^{(1)} = ||u_{mi}^{(1)} - f_{ni}^{(1)} ||^2 \); \( R_1 \) is the \( M_1 \times N \) label indicator matrix, the entry \( r_{mn}^{(1)} = 1 \) if the \( n \)th primitive \( v_n \) is labeled with the \( m \)th visual word \( u_{mi}^{(1)} \), otherwise.

The objective function of Eq. 1 can be solved by standard EM algorithm, i.e., iteratively updating \( R \) (E-step) and \( D \) (M-step). However, it suffers from several flaws. First, \( k \)-means clustering assumes all the visual primitives are independent from each other, and each potential cluster is convex in the feature space. Such assumptions are too strong to hold since primitives have spatial dependencies with one another and the clusters may have arbitrary shapes. Second, there are also feature dependencies among different types of features, which are discarded in the traditional \( k \)-means clustering method.

As a result, we should take into account multi-view features and spatial relations of primitives. In order to investigate the local spatial relations of primitives, we focus on the local spatial neighborhoods (e.g., \( k \)-NN or \( \epsilon \)-NN) of each instance \( v_i \), denoted as \( \{v_{i_1}, v_{i_2}, \ldots, v_{i_{K(i)}}\} \), where \( K(i) \) is an integer measured the size ( \( K(i) + 1 \) ) of this neighborhood.

Fig. 1 illustrates our idea. Without loss of generality, as it shows, each instance \( v_i \) has three types of features (i.e., \( c = 3 \)), where each type of feature can produce a visual word lexicon by \( k \)-means clustering. For example, \( \{A, B, C, \ldots\} \) denotes one type of visual word lexicon. Each visual word, i.e., \( A, B, C, \ldots, a, b, c, \ldots, \alpha, \beta, \gamma, \ldots \), can be represented as a prototype \( u_{mi}^{(j)} \) \( \in \mathbb{R}^{i_1} \). Combining such different types of visual words, we can partition the visual primitives into feature patterns: \( \{\square, \vee, \times, \ast, \cdot, \dagger, \nabla, \ast, \bullet, \ldots\} \). This partition will be more accurate and robust than the one that use a single type of visual word lexicon. Furthermore, because there are spatial dependencies among visual primitives, we can explore the co-occurrences of them. Hence, we group visual primitives into more meaningful spatial patterns in a higher level.

As a matter of fact, spatial patterns capture the spatial dependencies among feature patterns (note that visual primitives are clustered or labeled into feature patterns). And analogously, feature dependencies among multiple types of visual words are captured by feature patterns. These dependencies essentially provide the contextual information among the given visual primitives. They can be used to resolve the ambiguities and uncertainties in the low-level cluster representations (feature patterns and visual words). In specific, spatial patterns can help to tune feature patterns that of ambiguity and uncertainty, afterwards, all the visual words will also be adjusted due to the tuned feature patterns. Then several new visual words can learn more accurate feature
patterns and spatial patterns from bottom to up again. This is a self-learning process between low level features and high level patterns. Such an approach will for visual primitives clustering results (feature patterns), and will produce more meaningful patterns (spatial patterns).

Now, we would like to interpret the dependencies among different types of features with respect to the visual primitive $v_i$ as feature context of $v_i$, the dependencies among feature patterns with respect to $v_i$ as spatial context of $v_i$, and give the following formulations and definitions.

Suppose the dataset $D_v$ potentially have $M_v$ visual words from the $j$th type of features $\{f_l^{(j)}\}_{i=1}^N$, $M$ feature patterns, and $M_s$ spatial patterns. We then can represent feature contexts in a matrix form:

$$T_v = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_e \end{bmatrix},$$

(2)

where $R_j$ is the $M_j \times N$ binary label indicator matrix obtained from features $\{f_l^{(j)}\}_{i=1}^N$, the entry $r_{jn}$ = 1 only if $v_n$ is labeled with the $n$th discovered visual word; $T_v$ has size of $\sum_j M_j \times N$; and the $i$th column of $T_v$ is just the feature context of $v_i$, denoted as $t_i^{(v)}$.

**Definition 1 (Feature context).** The feature context of the visual primitive $v_i$ refers to the co-occurrence of multi-view visual words that $v_i$ belongs to.

We denote the local spatial neighbor relations of the dataset $D_v$ as a $N \times N$ binary matrix $Q_s$, where the entry $q_{ij}$ = 1 only if $v_i$ and $v_j$ are local spatial neighbors. According to [1], the spatial contexts in a matrix form can be represented as:

$$T_s = R_v Q_s,$$

(3)

where $R_v$ is the $M \times N$ binary label indicator matrix obtained for feature contexts; the entry $r_{vn}$ = 1 only if $v_n$ is labeled with the $n$th discovered feature patterns; $T_s$ has size of $M \times N$; and the $i$th column of $T_s$ is just the spatial context of $v_i$, denoted as $t_i^{(s)}$.

**Definition 2 (Spatial context).** The spatial context of the visual primitive $v_i$ refers to the collocated feature patterns appearing in the local spatial neighborhoods of $v_i$ in the original image space.

Our expectation is to use the correlations among visual primitives, feature contexts, and spatial contexts to make visual primitives be clustered into the correct visual words, feature contexts be clustered into the correct feature patterns, and spatial contexts be clustered into the correct spatial patterns. In order to reach this purpose, we propose a regularized $k$-means clustering:

$$J = \sum_{j=1}^e \sum_{m=1}^{M_j} \sum_{n=1}^N r_{mn}^{(j)} d_{n}(u_n^{(j)}, f_j^{(j)}) + \lambda_v \sum_{m=1}^M \sum_{n=1}^N r_{mn}^{(v)} (u_n^{(v)}, t_n^{(v)}) + \lambda_s \sum_{m=1}^M \sum_{n=1}^N r_{mn}^{(s)} (u_n^{(s)}, t_n^{(s)})$$

\[ = \sum_{i=1}^e \left( tr(R_i^T D_i) + \lambda_v tr(R_i^T D_v) + \lambda_s tr(R_i^T D_s) \right) \]

$$J_1 + J_2 + J_3,$$

(4)

where

- $\lambda_v > 0$ and $\lambda_s > 0$ are parameters for regularization;
- $R_v$ is the $M_v \times N$ binary label indicator matrix obtained for spatial contexts $\{t_i^{(v)}\}_{i=1}^N$, the entry $r_{mn}^{(v)} = 1$ only if $v_n$ is included in the $n$th discovered spatial patterns $u_n^{(s)}$ (with size of $M x 1$);
- $D_j$ is the $M_j \times N$ distortion matrix obtained from $j$th type of features $\{f_l^{(j)}\}_{i=1}^N$, the entry $d_{mn}^{(j)} = d_j(u_n^{(j)}, f_l^{(j)})$, i.e., the distortion between $u_n^{(j)}$ and the $m$th visual word $f_l^{(j)}$ (with size of $l_j \times 1$);
- $D_v$ is the $M_v \times N$ distortion matrix obtained from feature contexts $\{t_i^{(v)}\}_{i=1}^N$, the entry $d_{mn}^{(v)} = d_v(u_n^{(v)}, t_n^{(v)})$, i.e., the distortion between $t_n^{(v)}$ and the $m$th feature pattern $u_n^{(v)}$ (with size of $\sum_j M_j \times 1$);
- $D_s$ is the $M_s \times N$ distortion matrix obtained from spatial contexts, with the entry $d_{mn}^{(s)} = d_s(u_n^{(s)}, t_n^{(s)})$, i.e., the distortion between $t_n^{(s)}$ and the $m$th spatial pattern $u_n^{(s)}$;
- $R$ and $D$ are $\sum_j M_j \times cN$ block diagonal matrices from $\{R_i\}_{i=1}^e$ and $\{D_i\}_{i=1}^e$, respectively.
- $J_1, J_2,$ and $J_3$ correspond to the quantization distortions of multi-view features, feature contexts, and spatial contexts.

This seemingly simple objective function of Eq. 4 is actually challenging. We are not allowed to minimize $J$ by minimizing $J_1, J_2,$ and $J_3$ separately because they are correlated among each other. We will in Sec. IV show how to decouple the dependencies among them and propose our algorithm to solve this optimization function. Some major notations are listed in Table I.

**IV. OUR SOLUTION**

**A. Formulation**

Suppose $f_i^{(j)} \in \mathbb{R}^d$, the distortion matrix $D_j$ can be naturally measured by Euclidean square distance, i.e., for $j = 1, 2, \cdots, N$, we have
have to decouple the dependencies among the items of Eq. 4.

In a similar manner, we use the Hamming-like distortions. But for the same reason, it is equivalent to

Table I
Notations of some major symbols.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Size</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_j$</td>
<td>$M_j \times N$</td>
<td>binary label indicator matrix for the $j$th type of features ${ f^{(j)}<em>{i} }</em>{i=1}^{N}$; the entry $r^{(j)}<em>{mn}$ labels 1 only if the feature $f^{(j)}</em>{m}$ is clustered into the $mn$th visual word $u_{mn}^{(j)}$;</td>
</tr>
<tr>
<td>$U_j$</td>
<td>$I_j \times M_j$</td>
<td>visual word matrix from the $j$th type of features ${ f^{(j)}<em>{i} }</em>{i=1}^{N}$; the $r$th column is the visual word $u^{(j)}_r$;</td>
</tr>
<tr>
<td>$D_j$</td>
<td>$M_j \times N$</td>
<td>distortion matrix from the $j$th type of features ${ f^{(j)}<em>{i} }</em>{i=1}^{N}$; the entry $d^{(j)}_{mn}$ refers to the distortion between the visual word $u^{(j)}_m$ and the feature vector $f^{(j)}_n$;</td>
</tr>
<tr>
<td>$R_w$</td>
<td>$M \times N$</td>
<td>binary label indicator matrix for feature contexts ${ t^{(w)}<em>{i} }</em>{i=1}^{N}$; the entry $r^{(w)}_{mn}$ labels 1 only if the feature context $t^{(w)}<em>m$ is clustered into the $mn$th feature pattern $u^{(w)}</em>{mn}$;</td>
</tr>
<tr>
<td>$U_w$</td>
<td>$\sum_{j=1}^{J} M_j \times M$</td>
<td>feature pattern matrix; the $r$th column is the feature pattern $u^{(w)}_r$;</td>
</tr>
<tr>
<td>$D_w$</td>
<td>$M \times N$</td>
<td>distortion matrix from feature contexts ${ t^{(w)}<em>{i} }</em>{i=1}^{N}$; the entry $d^{(w)}_{mn}$ refers to the distortion between the feature pattern $u^{(w)}_m$ and the feature context $t^{(w)}_n$;</td>
</tr>
<tr>
<td>$R_s$</td>
<td>$M_s \times N$</td>
<td>binary label indicator matrix for spatial contexts ${ t^{(s)}<em>{i} }</em>{i=1}^{N}$; the entry $r^{(s)}_{mn}$ labels 1 only if the spatial context $t^{(s)}<em>m$ is clustered into the $mn$th spatial pattern $u^{(s)}</em>{mn}$;</td>
</tr>
<tr>
<td>$U_s$</td>
<td>$M \times M_s$</td>
<td>spatial pattern matrix, the $r$th column is the spatial pattern $u^{(s)}_r$;</td>
</tr>
<tr>
<td>$D_s$</td>
<td>$M_s \times N$</td>
<td>distortion matrix from spatial contexts ${ t^{(s)}<em>{i} }</em>{i=1}^{N}$; the entry $d^{(s)}_{mn}$ refers to the distortion between the spatial pattern $u^{(s)}_m$ and the spatial context $t^{(s)}_n$;</td>
</tr>
</tbody>
</table>

\[
d^{(j)}_{mn} = d_j \left( u^{(j)}_m, f^{(j)}_n \right) = \left\| u^{(j)}_m - f^{(j)}_n \right\|^2 . \tag{5}
\]

If we denote $Q_v$ as a $cN \times N$ matrix which is concatenated from $c$ identity matrices of size $N \times N$, we have

\[
T_v = RQ_v , \tag{6}
\]

where $Q_v$ is called the feature neighbor relation matrix.

Since $T_v$ is binary, we will use Hamming distance to represent $D_v$. But for the same reason, it is equivalent to Euclidean square distortion. $D_v$ can be represented by

\[
D_v = -2U_v^T T_v + 1_T T_v + U_v^T 1_{U_v} = -2U_v^T RQ_v + 1_T RQ_v + U_v^T 1_{U_v}, \tag{7}
\]

where $1_{T_v}$ is $M \times \sum_{j=1}^{J} M_j$ all 1 matrix; and $1_{U_v}$ is $\sum_{j=1}^{J} M_j \times N$ all 1 matrix.

In a similar manner, we use the Hamming-like distortions to formulate $D_s$ as

\[
D_s = -2U_s^T T_s + 1_T T_s + U_s^T 1_{U_s} = -2U_s^T R_s Q_s + 1_T R_s Q_s + U_s^T 1_{U_s}, \tag{8}
\]

where $1_{T_s}$ is the $M_s \times M$ all 1 matrix, and $1_{U_s}$ is the $M \times M_s$ all 1 matrix.

B. A Bottom-up/Top-down Iterative Refinement

We cannot estimate the unknown variables simultaneously in Eq. 4, since there are correlations among them. So we have to decouple the dependencies among the items of Eq. 4.

Above all, we initialize visual words, feature patterns and spatial patterns gradually by $k$-means, which are shown in Fig. 2 as "Initialization".

Next, we can take each of $R_s$, $R$, and $R_w$, as the common factor for extraction. We derive Eq. 4 as

\[
J(R, R_w, R_s, D, D_w, D_s) = tr(R^T H) \quad + tr(R^T D) \quad + \lambda_s tr(R_s^T U_s^T 1_{U_s}) \quad \tag{9}
\]

\[
\quad + tr(R^T H) \quad + \lambda_s tr(R_s^T U_s^T 1_{U_s}) \quad \tag{10}
\]

\[
\quad + tr(R^T H) \quad + \lambda_s tr(R_s^T U_s^T 1_{U_s}) \quad \tag{11}
\]

where

\[
\lambda_s = \lambda_s D_s - \lambda_s (2U_s^T - 1_{T_s})^T R_s Q_s^T , \tag{12}
\]

\[
H \quad = D - \lambda_s (2U_s^T - 1_{T_s})^T R_s Q_s^T , \tag{13}
\]

\[
H_w \quad = \lambda_s D_w , \tag{14}
\]

where the size of $H_w$, $H$ and $H_s$ are $M \times N$, $\sum_{j=1}^{J} M_j \times cN$ and $M_s \times N$, respectively; and $H$ contains $c$ blocks $\{H_j\}_{j=1}^{c}$.

We can successively update the three label indicator matrices $R_w$, $R$, and $R_s$ when the cluster centroid matrices $U_v$, $U_s$ ( $\sum_{j=1}^{J} d_j \times \sum_{j=1}^{c} M_j$ block diagonal matrix from $\{U_{j}^{(w)}\}_{j=1}^{c}$), and $U_s$ are given. According to the minimization requirement of the objective function defined by Eq. 4, the following label indicator matrices update criterion will be adopted, $\forall n = 1, 2, \cdots, N$,

\[
r^{(w)}_{mn} = \begin{cases} 1 & m = \arg \min_k h^{(w)}_{kn} \, , \\ 0 & \text{otherwise} \end{cases} , \tag{15}
\]

\[
r^{(j)}_{mn} = \begin{cases} 1 & m = \arg \min_k h^{(j)}_{kn} \, , \\ 0 & \text{otherwise} \end{cases} , \tag{16}
\]

\[
r^{(s)}_{mn} = \begin{cases} 1 & m = \arg \min_k h^{(s)}_{kn} \, , \\ 0 & \text{otherwise} \end{cases} . \tag{17}
\]

The derivation is shown in Appendix.

Although spatial contexts are not necessarily binary, the distortion derived from Hamming can also be perceived as a penalty measure, which is called Hamming-like distortion.
the objective function defined by Eq. 4 decreases. We show this in Th. 1. The objective function defined by Eq. 4 is monotonic decreasing at each step. We summarize this in Alg. 1. Based on the above discussions, we propose an iterative algorithm for contextual clustering described in Alg. 1. This Algorithm is convergent since the solution space of $R$, $R_v$, and $R_s$ are discrete and finite, and the objective function defined by Eq. 4 is monotonic decreasing at each step. We summarize this in Th. 1.

Algorithm 1: Contextual Clustering Algorithm

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>${v_i}<em>{i=1}^N$: feature neighbor relations: $Q_v$; spatial neighbor relations: $Q_s$; parameters: ${M_j}</em>{j=1}^c$, $M_v$, $\lambda_v$, $\lambda_r$</td>
<td>cluster centroids ${U_i}<em>{i=1}^c$, $U_v$, $U_s$; cluster indicator labels ${R_i}</em>{i=1}^c$, $R_v$, $R_s$</td>
</tr>
</tbody>
</table>

1. **init**: perform $k$-means clustering from bottom to up to obtain $\{U_i\}_{i=1}^c$, $U_v$, $U_s$;
2. while $J$ is decreasing do
   3. **R-step**: fix $\{U_i\}_{i=1}^c$, $U_v$, $U_s$, iteratively top-down/bottom-up update $\{R_i\}_{i=1}^c$, $R_v$, $R_s$;
   4. if $J$ is decreasing then goto **R-step**;
   5. else goto **D-step**;
   6. **D-step**: fix $\{R_i\}_{i=1}^c$, $R_v$, $R_s$, update $\{U_i\}_{i=1}^c$,
   7. return $\{U_i\}_{i=2}^c$, $U_v$, $U_s$, $\{R_i\}_{i=1}^c$, $R_v$, $R_s$.

**Theorem 1 (Convergence)**. The proposed contextual clustering algorithm in Alg. 1 can converge in finite steps.

where $h_{kn}^{(c)}$, $r_{kn}^{(v)}$, $h_{kn}^{(j)}$, $r_{kn}^{(s)}$ and $r_{mn}^{(v)}$ are the entries of $H_v$, $R_v$, $H_j$, $R_j$, $H_s$ and $R_s$ respectively. This process is demonstrated in Fig. 2 at “R-step”, where as long as the objective function $R$ defined by Eq. 4 is decreasing, $R_v$ and $R$ will be continually tuned from high level to low level, followed by the bottom-up update of $R_v$ and $R_s$.

Furthermore, provided the label indicator matrices $R_v$, $R$, and $R_s$, the corresponding centroid matrices $U_v$, $\{U_j\}_{j=1}^c$, and $U_s$ can be updated, and so as the corresponding distortion matrices $D_v$, $\{D_j\}_{j=1}^c$, and $D_s$, which will also make the objective function defined by Eq. 4 decrease. We show this procedure in Fig. 2 at “D-step”.

Based on the above discussions, we propose an iterative algorithm for contextual clustering described in Alg. 1. This Algorithm is convergent since the solution space of $R$, $R_v$, and $R_s$ are discrete and finite, and the objective function defined by Eq. 4 is monotonic decreasing at each step. We summarize this in Th. 1.

**Theorem 1 (Convergence)**. The proposed contextual clustering algorithm in Alg. 1 can converge in finite steps.
In our contextual clustering, we can set the parameter 
\( \lambda_s = 0 \) to consider feature contexts only. Besides, we can remove the feature context term \( J_2 \) (i.e., let \( \lambda_r = 0 \)) from objective function of Eq. 4 when we don’t consider multiple types of features. In [1], these two special cases have been investigated separately.

V. Experiments

To evaluate our approach, we perform clustering on individual images in the first experiment. For each key patch, we employ two feature views: SIFT features [3] and Color Histograms (CH) [16]. To find co-occurred spatial patterns in images, the spatial relations of different key patches are also used to provide contextual information.

In the second experiment, we perform region clustering on the dataset of MSRC-V2 [4]. The ground-truth labeling of this dataset is provided by [17]. To describe each region, we employ three views as in [16]: Texton Histograms (TH), Color Histograms (CH), and pyramid of HOG (pHOG) [18]. Each region category is one potential spatial pattern and we want to cluster image regions into their corresponding region categories.

The dimension of TH, CH, pHOG, and SIFT is 400, 69, 680 and 128, respectively. All the features are normalized by Euclidean distance. Also for simplicity, we set \( M_j = M, \forall j = 1, 2, \cdots, c \), and let \( \lambda_r = \tau_v \| J_i^v / J_i^{\tau_v} \|, \lambda_s = \tau_s \| J_i^s / J_i^{\tau_s} \| \), where \( J_i^j \) is the initial value of \( J_i \) defined by Eq. 4. The nonnegative constants \( \tau_v \) and \( \tau_s \) are the auxiliary parameters to balance the distortions among the levels of visual words, feature patterns, and spatial patterns.

A. Image Pattern Discovery

1) Texture Image: We test the proposed method on texture image modelling [19], [20], [21]. Fig. 3 shows the synthetical texture image with four visual patterns. First, we extract 3052 SIFT patches [3] which are described by SIFT features [3] and Color Histograms (CH) [16] (CH). Second, we use the \( K \)-nearest neighbors as the spatial neighbors with \( K = 8 \) in our contextual clustering, and let \( M = 10, M_s = 4, \tau_v = 0.85, \tau_s = 0.5 \). The 1st column of Fig. 4 shows the spatial pattern clusters discovered using our method. It can be seen that they are corresponding to the real patterns in the image. In contrast, the 2nd column of Fig. 4 shows the discovered clusters when only use SIFT features and spatial contextual information [1] (the same condition with our approach except absence of the parameter \( \tau_v \)). It should be noticed that in the latter method, there is much confusion between face pattern and edge pattern. By contrast, our method gives more accurate patterns. Fig. 8(b) depicts the convergent process of our method.

B. Image Region Clustering

1) Clear Spatial Contextual Relation: We select a collection of images with two region pairs that often appear together in an image: “sheep+grass” and “bicycle+road” shown in Fig. 7, where 122 regions are included. Here, the spatial neighbors of a region is the regions occurring in this spatial pattern and also stem from the same spatial contextual relations. In this experiment, three types of features for regions are involved [16]: TH, CH and pHOG, and the following region clustering are carried out:

- cluster regions by each type of features using traditional \( k \)-means; (\( k \)-means)
- concatenate the three types of features for traditional \( k \)-means clustering; (\( TH+CH+PHOG, k \)-means)

2) Real Image: We also perform clustering on real images, and present an example in Fig. 6, where 422 SIFT key patches are extracted. In our clustering, the patches are represented by two features of SIFT and CH, spatial neighbors are defined by \( K \)-nearest neighbors with \( K = 12 \), and other parameters are set as \( M = 10, M_s = 3, \tau_v = 1.5, \tau_s = 0.8 \). The discovered spatial patterns are shown in the 1st column of Fig. 6, which are corresponding to three real patterns in the image: face, logo, and edge. The 2nd column of Fig. 6 shows the detected common patterns using only SIFT features and spatial contextual information (the same condition with our method except absence of the parameter \( \tau_v \)). It should be noticed that in the latter method, there is much confusion between face pattern and edge pattern. By contrast, our method gives more accurate patterns. Fig. 8(b) depicts the convergent process of our method.

Figure 3. An image with four visual patterns: purple circle inlaid with white petals, blue star, pink diamond inlaid with white star, and blue diamond inlaid with white star (best viewed in color).
Figure 4. Results of spatial pattern discovery on the image shown in Fig. 3, where we use red/cyan to indicate the members of each cluster. The 1st column show the results of our proposed method, i.e., clustering using multi-view features (SIFT + CH) and spatial contexts; and the 2nd column show the results of [1], i.e., clustering using SIFT features and spatial contexts (best viewed in color).

Figure 5. Each detected spatial pattern consists of a set of bases, where each base of a spatial pattern cluster (Fig. 4(a)) is obtained by taking average of patches that occur in this spatial pattern cluster and also come from the same feature pattern. The images are re-scaled and normalized for display (best viewed in color).

- perform clustering only using spatial contextual information on each type of features [1] (spatial context);
- perform clustering only using feature contextual information [1] (feature context);
- perform clustering using both information of feature contexts and spatial contexts (proposed method, feature context + spatial context).

Figure 6. Results of spatial pattern discovery on a group photo, where we use green/cyan to indicate the members of each cluster. The 1st column show the results of our proposed method, i.e., clustering using multi-view features (SIFT + CH) and spatial contexts; and the 2nd column show the results of [1], i.e., clustering using SIFT features and spatial contexts (best viewed in color).

Figure 7. Images with two region pairs of “sheep+grass” and “bicycle+road”. Each pair contains 27 images, the region numbers are: sheep 31, grass 32, bicycle 27, and road 32.

The results are listed in Table II, where we can see the spatial contextual information cannot improve the clustering performance significantly when the $k$-means clustering results is poor. However, there is also a considerable
method[

<table>
<thead>
<tr>
<th>feature</th>
<th>method</th>
<th>$k$</th>
<th>parameters</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH</td>
<td>$k$-means</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>53.43%</td>
</tr>
<tr>
<td>TH</td>
<td>spatial context</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>27.05%</td>
</tr>
<tr>
<td>CH</td>
<td>$k$-means</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>41.80%</td>
</tr>
<tr>
<td>CH</td>
<td>spatial context</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>40.98%</td>
</tr>
<tr>
<td>pHOG</td>
<td>$k$-means</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>37.30%</td>
</tr>
<tr>
<td>pHOG</td>
<td>spatial context</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>36.88%</td>
</tr>
<tr>
<td>TH+CH+pHOG</td>
<td>$k$-means</td>
<td>$k=4$</td>
<td>$M=4, M_s=2, \tau_s=7$</td>
<td>40.98%</td>
</tr>
<tr>
<td>TH+CH+pHOG</td>
<td>spatial context</td>
<td>$M=4, M_s=4, \tau_v=10$</td>
<td>40.16%</td>
<td></td>
</tr>
<tr>
<td>TH+CH+pHOG</td>
<td>feature context</td>
<td>$M=4, M_s=4, \tau_v=10$</td>
<td>11.48%</td>
<td></td>
</tr>
<tr>
<td>TH+CH+pHOG</td>
<td>feature + spatial context (proposed)</td>
<td>$M=4, M_s=2, \tau_v=10, \tau_s=7$</td>
<td>6.56%</td>
<td></td>
</tr>
</tbody>
</table>

improvement when a better $k$-means clustering result is obtained. Especially, imposing spatial contextual information on TH, the accuracy is improved by 7.38% (error decreases from 34.43% to 27.05%), using both information of spatial contexts and feature contexts (i.e., our method, with error 6.56%) is also much better than using only feature contextual information (with error 11.48%). Fig. 8(c) shows the convergent process of our proposed method. The parameters are given in Table II.

2) Unobvious Spatial Contextual Relation: We also test our method on ambiguous spatial contextual information. In this experiment, 150 test images with their region segmentations are collected. From Table III shows the list of the collection. We can see that the spatial contextual information are not clear enough, and different image regions may have the same neighbor(s) (e.g., “grass” occurs in three different scenes).

Same with Sec. V-B1, each region has three types of features: TH, CH, pHoG, and the spatial neighbors of a region is defined by the regions occurring in the same image. The region clustering results of $k$-means clustering, feature contextual clustering, and our proposed clustering are given in Table IV, where, for each region category, 10% seeds from ground-truth are randomly chosen for initialization. We can see that the proposed feature contextual clustering performance much better than each type of features or concatenated features for $k$-means clustering. Moreover, although some different image regions have the same neighbor(s), the spatial relations among regions also provide useful information for clustering, which can be seen from that our proposed method (feature context + spatial context) outperforms the feature contextual method (error decreases by 1.89%). The convergent process is shown in Fig. 8(d). We set the parameters as: $k=9$ for $k$-means clustering; $\tau_v=0.7, M=9, M_v=9$ for feature contextual clustering; $\tau_v=0.7, \tau_s=0.01, M=9, M_s=5$ for our proposed clustering.

<table>
<thead>
<tr>
<th>segmentation of image</th>
<th>number of images</th>
</tr>
</thead>
<tbody>
<tr>
<td>sheep+grass</td>
<td>30</td>
</tr>
<tr>
<td>cow+grass</td>
<td>30</td>
</tr>
<tr>
<td>aeroplane+grass+sky</td>
<td>30</td>
</tr>
<tr>
<td>boat+water</td>
<td>30</td>
</tr>
<tr>
<td>bicycle+road</td>
<td>30</td>
</tr>
</tbody>
</table>

Table IV

Comparison among $k$-means clustering and two kinds of contextual clustering for image regions from images listed in Table III. For each region category, 10% seeds from ground-truth are randomly chosen for initialization.

VI. CONCLUSION
Because of the spatial structures and content variations of complex visual patterns, they greatly challenge most existing data mining methods to discover meaningful visual patterns in images. We propose a novel pattern discovery method to construct low-level visual primitives, e.g. local image patches or regions, into high-level visual patterns of spatial structures. Instead of ignoring the spatial dependency among visual primitives and simply performing $k$-means clustering to obtain the visual vocabulary, we explore both feature and spatial contexts and discover the co-occurrence patterns to resolve the ambiguities among visual primitives. To solve the regularized $k$-means clustering, an iterative top-down/bottom-up procedure is developed. It performs self-learning to iteratively refine the pattern discovery results and guarantees to converge. Compared to our previous work [1], as both spatial and feature contexts are utilized, a better clustering result can be obtained. The experiments on image
texton discovery and image region clustering justify the advantages of the proposed method.

ACKNOWLEDGMENT

This work was supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

APPENDIX

A. Matrix form of Hamming distortion

The Hamming distortion between two \( n \times 1 \) binary vectors \( \mathbf{x}_i \) and \( \mathbf{y}_j \) can be represented as

\[
d(\mathbf{x}_i, \mathbf{y}_j) = n - [x_i^T y_j + (1 - x_i)^T (1 - y_j)]
\]

where \( \mathbf{1} \) is the \( n \times 1 \) all 1 vector.

Suppose that we have \( \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \) and \( \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_j] \), and the \( i \)th row and \( j \)th column element of the \( s \times t \) matrix \( \mathbf{D} \) denote the Hamming distortion between \( \mathbf{x}_i \) and \( \mathbf{y}_j \), i.e., \( d(\mathbf{x}_i, \mathbf{y}_j) \). The matrix form of Hamming distortion is given by

\[
\mathbf{D} = -2\mathbf{X}^T \mathbf{Y} + \mathbf{1}_Y \mathbf{Y}^T + \mathbf{X}^T \mathbf{1}_X,
\]

where \( \mathbf{1}_Y \) is \( s \times n \) all 1 matrix, and \( \mathbf{1}_X \) is \( n \times t \) all 1 matrix.

B. Derivations of Eqs. 9, 10, 11

We first show how to obtain Eq. 9 from Eq. 4. Substitute Eq. 8 into Eq. 4, and apply the three properties of matrix trace:

\[
\text{tr} (\mathbf{A}) = \text{tr} (\mathbf{A}^T) \quad (\text{Eq. 19})
\]

for any square matrices \( \mathbf{A} \):

\[
\text{tr} (\mathbf{AB}) = \text{tr} (\mathbf{BA}) \quad (\text{Eq. 20})
\]

for \( m \times n \) matrix \( \mathbf{A} \) and \( n \times m \) matrix \( \mathbf{B} \); \n
\[
\text{tr} (\alpha \mathbf{A} + \beta \mathbf{B}) = \alpha \text{tr} (\mathbf{A}) + \beta \text{tr} (\mathbf{B}) \quad (\text{Eq. 21})
\]

for constants \( \alpha, \beta \), and square matrices \( \mathbf{A} \) and \( \mathbf{B} \). We then have the following derivation:

\[
\mathbf{J} = \text{tr} (\mathbf{R}^T \mathbf{D}) + \lambda_s \text{tr} (\mathbf{R}_s^T \mathbf{D}_s) + \lambda_s \text{tr} (\mathbf{R}_s^T \mathbf{D}_s)
\]

where \( \lambda_s \) is defined in Eq. 4, \( \mathbf{J}_{max} \) is the initial value of \( \mathbf{J} \) defined in Eq. 4 (for normalizing the curve of all the \( \mathbf{J}_i \) and \( \mathbf{J}_j \)), and error denotes the error rate of region clustering in the experiments of Sec. V-B.
Similar to the above derivation, we can obtain Eq. 10 by substituting Eq. 7 into Eq. 4:

\[
\mathbf{J} = \text{tr} \left( \mathbf{R}^T \mathbf{D} \mathbf{R} \right) + \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) + \lambda_v \text{tr} \left( \mathbf{R}^T_v \mathbf{D}_v \right)
\]

\[
= \text{tr} \left( \mathbf{R}^T \mathbf{D} \right) + \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) + \lambda_v \text{tr} \left( \mathbf{R}^T_v \mathbf{D}_v \right)
\]

\[
+ \lambda_s \text{tr} \left[ \mathbf{R}^T_v \mathbf{D}_v \left( -2 \mathbf{U}^T_v \mathbf{Q}_v \mathbf{R}_v + \mathbf{1}_T \mathbf{R}_v \mathbf{Q}_v + \mathbf{1}_T^T \mathbf{U}_v \right) \right]
\]

\[
= \text{tr} \left( \mathbf{R}^T \mathbf{D} \right) + \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) + \lambda_v \text{tr} \left[ \mathbf{R}^T_v \left( \mathbf{U}_v \mathbf{1}_u \right) \right]
\]

\[
+ \lambda_s \text{tr} \left[ \mathbf{R}^T_v \left( -2 \mathbf{U}^T_v \mathbf{Q}_v + \mathbf{1}_T \mathbf{R}_v \mathbf{Q}_v \right) \right]
\]

\[
= \text{tr} \left( \mathbf{R}^T \mathbf{D} \right) + \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) + \lambda_v \text{tr} \left[ \mathbf{R}^T_v \left( \mathbf{U}_v \mathbf{1}_u \right) \right]
\]

\[
+ \lambda_s \text{tr} \left[ \mathbf{Q}_v \mathbf{R}^T \left( -2 \mathbf{U}^T_v + \mathbf{1}_T \right) \mathbf{R}_v \mathbf{Q}_v \right]
\]

\[
= \text{tr} \left( \mathbf{R}^T \mathbf{D} \right) + \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) + \lambda_v \text{tr} \left[ \mathbf{R}^T_v \mathbf{Q}_v \mathbf{R}_v \mathbf{Q}_v \mathbf{R}^T \left( -2 \mathbf{U}^T_v + \mathbf{1}_T \right) \mathbf{R}_v \mathbf{Q}_v \right]
\]

\[
+ \lambda_s \text{tr} \left[ \mathbf{Q}_v \mathbf{R}^T \mathbf{D}_v \right] + \lambda_v \text{tr} \left[ \mathbf{R}^T_v \left( \mathbf{U}_v \mathbf{1}_u \right) \right]
\]

Finally, note that \( \lambda_s \text{tr} \left( \mathbf{R}^T_s \mathbf{D}_s \right) = \text{tr} \left[ \mathbf{R}^T_s \left( \lambda_s \mathbf{D}_s \right) \right] \), this implies Eq. 11 from Eq. 4.

\[\text{REFERENCES}\]


