

Chapter 2

Context-Aware Discovery of Visual Co-occurrence Patterns

Abstract Once images are decomposed into a number of visual primitives, it is of great interests to cluster these primitives into mid-level visual patterns. However, conventional clustering of visual primitives, e.g., bag-of-words, usually ignores the spatial context and multi-feature information among the visual primitives and thus cannot discover mid-level visual patterns of complex structure. To overcome this problem, we propose to consider both spatial and feature contexts among visual primitives for visual pattern discovery in this chapter. We formulate the pattern discovery task as a multi-context-aware clustering problem and propose a self-learning procedure to iteratively refine the result until it converges. By discovering both spatial co-occurrence patterns among visual primitives and feature co-occurrence patterns among different types of features, the proposed method can better address the ambiguities of visual primitives.

Keywords Co-occurrence pattern discovery • Visual disambiguity • Multi-context-aware clustering • k -means regularization • Self-learning optimization

2.1 Introduction

It has been a common practice to build a visual vocabulary for image analysis by visual primitive clustering. However, most existing clustering methods ignore the spatial structure among the visual primitives [7], thus bringing unsatisfactory results. For example, the popular k -means clustering of visual primitives can lead to synonymous visual words that overrepresent visual primitives, as well as polysemous visual words that bring large uncertainties and ambiguities in the representation [5, 6].

Since visual primitives are not independent to each other, to better address the visual polysemous and synonymous phenomena, the ambiguities and uncertainties of visual primitives can be partially resolved through analyzing their spatial contexts [12, 13], i.e., other primitives in the spatial neighborhood. Two visual primitives, although exhibit dissimilar visual features, may belong to the same pattern if they have the same spatial contexts. Even though they share similar features, they may not

belong to the same visual pattern if their spatial contexts are completely different. Besides the spatial dependencies among visual primitives, a visual pattern can exhibit certain feature dependencies among multiple types of features or attributes as well. Therefore, it is equally interesting in discovering spatial and feature co-occurrence patterns in image data so that we can leverage visual patterns to improve the clustering of visual primitives.

To address the above problem, we propose to consider spatial and feature contexts among visual primitives for pattern discovery. By discovering spatial co-occurrence patterns among visual primitives and feature co-occurrence patterns among different types of features, our method can effectively reduce the ambiguities of visual primitive clustering. We formulate the pattern discovery problem as multi-context-aware clustering, where spatial and feature contexts are served as constraints of k -means clustering to improve the pattern discovery results. A novel self-learning procedure is proposed to integrate visual pattern discovery into the process of visual primitive clustering. The proposed self-learning procedure is guaranteed to converge, and experiments on real images validate the effectiveness of our method.

2.2 Multi-context-aware Clustering

In multi-context-aware clustering, each visual primitive $x_n \in \mathcal{X}$ is characterized by V types of features: $\{\mathbf{f}_n^{(v)}\}_{v=1}^V$, where $\mathbf{f}_n^{(v)} \in \mathbb{R}^{d_v}$. These features of x_n correspond to a feature context group $\mathcal{G}_n^{(f)}$. Meanwhile, collocating with a visual primitive in a local spatial neighborhood, the inclusive visual primitives constitute the spatial contexts of the central one. For each visual primitive $x_n \in \mathcal{X}$, we denote by $\mathcal{G}_n^{(s)} = \{x_n, x_{n_1}, x_{n_2}, \dots, x_{n_K}\}$ its *spatial context group*, which can be built by K -nearest neighbors (K -NN) or ε -nearest neighbors (ε -NN).

2.2.1 Regularized k -means Formulation with Multiple Contexts

Each type of features $\{\mathbf{f}_n^{(v)}\}_{n=1}^N$ can produce a feature word lexicon Ω_v ($|\Omega_v| = M_v$) by k -means clustering with the objective function (2.1) minimized.

$$Q_v = \sum_{m=1}^{M_v} \sum_{n=1}^N r_{mn}^{(v)} d_v(\mathbf{u}_m^{(v)}, \mathbf{f}_n^{(v)}) = \text{tr}(\mathbf{R}_v \mathbf{D}_v), \quad (2.1)$$

where

- $\{\mathbf{u}_m^{(v)}\}_{m=1}^{M_v}$ denote M_v quantized feature words after clustering, and they together form a feature word matrix $\mathbf{U}_v \in \mathbb{R}^{d_v \times M_v}$;

- $\mathbf{R}_v \in \mathbb{R}^{M_v \times N}$ is a binary label indicator matrix, and the entry $r_{mn}^{(v)} = 1$ only if $\mathbf{f}_n^{(v)}$ is labeled with the m th discovered feature word $\mathbf{u}_m^{(v)}$ via clustering;
- $\mathbf{D}_v \in \mathbb{R}^{M_v \times N}$ denotes a distortion matrix, and the entry of its m th row and n th column is given by $d_v(\mathbf{u}_m^{(v)}, \mathbf{f}_n^{(v)})$, i.e., the distortion between $\mathbf{u}_m^{(v)}$ and $\mathbf{f}_n^{(v)}$.

To consider multiple types of features, we let each $x_n \in \mathcal{X}$ generate a feature context transaction $\mathbf{t}_n^{(f)} \in \mathbb{R}^{\sum_{v=1}^V M_v}$ to represent $\mathcal{G}_n^{(f)}$.

Definition 2.1 (*Feature context transaction*) The feature context transaction of the visual primitive x_n , denoted by $\mathbf{t}_n^{(f)}$, refers to the co-occurrences of multiple types of feature words in the feature context group of x_n .

Using label indicator matrices $\{\mathbf{R}_v\}_{v=1}^V$ obtained from k -means clustering on the V types of features, we can represent the feature context transaction database as a binary matrix

$$\mathbf{T}_f = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \vdots \\ \mathbf{R}_V \end{bmatrix}. \quad (2.2)$$

Therefore, $\mathbf{T}_f \in \mathbb{R}^{\sum_{v=1}^V M_v \times N}$, and $\mathbf{t}_n^{(f)}$ is in the n th column of \mathbf{T}_f . Similar to single feature clustering, we propose to minimize the objective function (2.3) to obtain a mid-level feature pattern lexicon Ψ_f ($|\Psi_f| = M_f$), which actually provide a partition to the given data in \mathcal{X} using multiple features.

$$Q_f = \sum_{m=1}^{M_f} \sum_{n=1}^N r_{mn}^{(f)} d_f(\mathbf{u}_m^{(f)}, \mathbf{t}_n^{(f)}) = \text{tr}(\mathbf{R}_f \mathbf{D}_f), \quad (2.3)$$

where

- $\{\mathbf{u}_m^{(f)}\}_{m=1}^{M_f}$ denote M_f quantized feature patterns after clustering, and they form a feature pattern matrix $\mathbf{U}_f \in \mathbb{R}^{\sum_{j=1}^V M_j \times M_f}$;
- $\mathbf{R}_f \in \mathbb{R}^{M_f \times N}$ is a binary label indicator matrix, and the entry $r_{mn}^{(f)} = 1$ only if v_n is included the m th discovered feature pattern $\mathbf{u}_m^{(f)}$ via clustering;
- $\mathbf{D}_f \in \mathbb{R}^{M_f \times N}$ denotes a distortion matrix, and the entry of its m th row and n th column is given by $d_f(\mathbf{u}_m^{(f)}, \mathbf{t}_n^{(f)})$, i.e., the distortion between $\mathbf{u}_m^{(f)}$ and $\mathbf{t}_n^{(f)}$.

Besides multi-feature information, we further explore the spatial dependencies among visual primitives and represent $\mathcal{G}_n^{(s)}$ as a spatial context transaction.

Definition 2.2 (*Spatial context transaction*) The spatial context transaction of the visual primitive x_n , denoted by $\mathbf{t}_n^{(s)}$, refers to the co-occurrences of different categories of visual primitives appearing in the spatial context group of x_n .

The spatial context transaction database can be represented as a sparse integer matrix $\mathbf{T}_s \in \mathbb{R}^{M_f \times N}$, where each column is a spatial context transaction $\mathbf{t}_n^{(s)} \in \mathbb{Z}^{M_f}$. The entry



Fig. 2.1 Pattern discovery along the *solid arrows* and visual disambiguity along the *dashed arrows*

$t_{mn}^{(s)} = c$ indicates that the n th transaction contains c visual primitives belonging to the m th category. Similarly, we can find a higher level spatial pattern lexicon Ψ_s ($|\Psi_s| = M_s$) by clustering on spatial context transactions. The minimization objective function is given by

$$Q_s = \sum_{m=1}^{M_s} \sum_{n=1}^N r_{mn}^{(s)} d_s(\mathbf{u}_m^{(s)}, \mathbf{t}_n^{(s)}) = \text{tr}(\mathbf{R}_s \mathbf{D}_s), \quad (2.4)$$

where

- $\{\mathbf{u}_m^{(s)}\}_{m=1}^{M_s}$ denote M_s quantized spatial patterns after clustering, and they form a spatial pattern matrix $\mathbf{U}_s \in \mathbb{R}^{M_f \times M_s}$;
- $\mathbf{R}_s \in \mathbb{R}^{M_s \times N}$ is a binary label indicator matrix, and the entry $r_{mn}^{(s)} = 1$ only if v_n is included the m th discovered spatial pattern $\mathbf{u}_m^{(s)}$ via clustering;
- $\mathbf{D}_s \in \mathbb{R}^{M_s \times N}$ denotes a distortion matrix, and the entry of its m th row and n th column is given by $d_s(\mathbf{u}_m^{(s)}, \mathbf{t}_n^{(s)})$, i.e., the distortion between $\mathbf{u}_m^{(s)}$ and $\mathbf{t}_n^{(s)}$.

After having spatial patterns, we aim to refine visual primitive clustering of uncertainty. Such a refinement should enable spatial patterns to help improve feature pattern constructions. Afterward, each type of feature words will also be adjusted due to the tuned feature patterns. Then, the multiple types of updated feature words can learn more accurate feature patterns and spatial patterns from bottom up again. We show the idea in Fig. 2.1. To achieve this objective, we propose to minimize (2.1) regularized by (2.3) and (2.4). The objective function thus becomes

$$\begin{aligned} Q &= \sum_{v=1}^V \text{tr}(\mathbf{R}_v^T \mathbf{D}_v) + \lambda_f \text{tr}(\mathbf{R}_f^T \mathbf{D}_f) + \lambda_s \text{tr}(\mathbf{R}_s^T \mathbf{D}_s) \\ &= \underbrace{\text{tr}(\mathbf{R}^T \mathbf{D})}_{Q_\alpha} + \underbrace{\lambda_f \text{tr}(\mathbf{R}_f^T \mathbf{D}_f)}_{Q_\beta} + \underbrace{\lambda_s \text{tr}(\mathbf{R}_s^T \mathbf{D}_s)}_{Q_\gamma}, \end{aligned} \quad (2.5)$$

where

- $\lambda_f > 0$ and $\lambda_s > 0$ are constants for regularization;
- Q_α , Q_β , and Q_γ are the total quantization distortions of multiple types of features, the quantization distortion of feature context transactions, and the quantization distortion of spatial context transactions, respectively.
- \mathbf{R} and \mathbf{D} are block diagonal matrices from $\{\mathbf{R}_i\}_{i=1}^V$ and $\{\mathbf{D}_i\}_{i=1}^V$.

As Q_α , Q_β , and Q_γ are correlated among each other, it is intractable to minimize Q by minimizing the three terms separately, which makes the objective function of (2.5) a challenge. We will in Sect. 2.2.2 show how to decouple the dependencies among them and propose our algorithm to solve this optimization function.

2.2.2 Self-learning Optimization

We initialize feature words, feature patterns, and spatial patterns gradually by k -means clustering by minimizing (2.1), (2.3), and (2.4). During k -means clustering, we use squared Euclidean distance to measure $d_v(\cdot, \cdot)$ in each feature space. Since feature context transactions are binary, we use Hamming distance to measure $d_f(\cdot, \cdot)$, which leads to

$$\begin{aligned} \mathbf{D}_f &= -2\mathbf{U}_f^T \mathbf{T}_f + \mathbf{1}_{T_f} \mathbf{T}_f + \mathbf{U}_f^T \mathbf{1}_{U_f} \\ &= -2\mathbf{U}_f^T \mathfrak{R} \mathbf{Z}_f + \mathbf{1}_{T_f} \mathfrak{R} \mathbf{Z}_f + \mathbf{U}_f^T \mathbf{1}_{U_f}, \end{aligned} \quad (2.6)$$

where $\mathbf{1}_{T_f}$ is an $M \times \sum_{i=1}^V M_v$ all 1 matrix; $\mathbf{1}_{U_f}$ is a $\sum_{i=1}^V M_v \times N$ all 1 matrix; and $\mathbf{Z}_f \in \mathbb{R}^{V \times N}$ is the concatenation of V identity matrices of $N \times N$. Following (2.6), we can have a similar distortion matrix to spatial context transactions

$$\begin{aligned} \mathbf{D}_s &= -2\mathbf{U}_s^T \mathbf{T}_s + \mathbf{1}_{T_s} \mathbf{T}_s + \mathbf{U}_s^T \mathbf{1}_{U_s} \\ &= -2\mathbf{U}_s^T \mathbf{R}_f \mathbf{Z}_s + \mathbf{1}_{T_s} \mathbf{R}_f \mathbf{Z}_s + \mathbf{U}_s^T \mathbf{1}_{U_s}, \end{aligned} \quad (2.7)$$

where $\mathbf{1}_{T_s}$ is an $M_s \times M$ all 1 matrix; $\mathbf{1}_{U_s}$ is an $M \times N$ all 1 matrix; and \mathbf{Z}_s is an $N \times N$ matrix, whose entry $q_{ij} = 1$ only if x_i and x_j are local spatial neighbors. It is worth noting that the matrix (2.7) no longer indicates pairwise distances but only distortion penalties, unless spatial context transactions are all binary.

To decouple the dependencies among the terms of (2.5), we take each of \mathbf{R}_f , \mathbf{R} , and \mathbf{R}_s as the common factor for extraction and derive (2.5) as:

$$\begin{aligned} &Q(\mathfrak{R}, \mathbf{R}_f, \mathbf{R}_s, \mathfrak{D}, \mathbf{D}_f, \mathbf{D}_s) \\ &= tr(\mathbf{R}_f^T \mathbf{H}_f) + tr(\mathfrak{R}^T \mathfrak{D}) + \lambda_s tr(\mathbf{R}_s^T \mathbf{U}_s^T \mathbf{1}_{U_s}) \end{aligned} \quad (2.8)$$

$$= tr(\mathfrak{R}^T \mathfrak{H}) + \lambda_s tr(\mathbf{R}_s^T \mathbf{D}_s) + \lambda_f tr(\mathbf{R}_f^T \mathbf{U}_f^T \mathbf{1}_{U_f}) \quad (2.9)$$

$$= tr(\mathbf{R}_s^T \mathbf{H}_s) + tr(\mathfrak{R}^T \mathfrak{D}) + \lambda_f tr(\mathbf{R}_f^T \mathbf{D}_f), \quad (2.10)$$

in which

$$\mathbf{H}_f = \lambda_f \mathbf{D}_f - \lambda_s (2\mathbf{U}_s^T - \mathbf{1}_{T_s})^T \mathbf{R}_s \mathbf{Z}_s^T, \quad (2.11)$$

$$\mathfrak{H} = \mathfrak{D} - \lambda_f (2\mathbf{U}_f^T - \mathbf{1}_{T_f})^T \mathbf{R}_f \mathbf{Z}_f^T, \quad (2.12)$$

$$\mathbf{H}_s = \lambda_s \mathbf{D}_s, \quad (2.13)$$

Algorithm 1: Visual Pattern Discovery with Multi-Context-Aware Clustering (MCAC)

Input: $\mathcal{X} = \{x_n\}_{n=1}^N$; \mathbf{Z}_f ; \mathbf{Z}_s ; parameters: $\{M_v\}_{v=1}^V$, M_f , M_s , λ_f , λ_s
Output: feature word lexicons: $\{\Omega_v\}_{v=1}^V$ ($\{\mathbf{U}_v\}_{v=1}^V$); feature pattern lexicon: $\Psi_f(\mathbf{U}_f)$; spatial pattern lexicon: $\Psi_s(\mathbf{U}_s)$; clustering results $\{\mathbf{R}_v\}_{v=1}^V$, \mathbf{R}_f , \mathbf{R}_s
 // Initialization
 1: perform k -means clustering from bottom up to obtain $\{\mathbf{U}_i\}_{i=1}^V$, \mathbf{U}_f , \mathbf{U}_s
 // Main loop
 2: **repeat**
 3: **repeat**
 4: **R-step:** fix $\{\mathbf{U}_i\}_{i=1}^V$, \mathbf{U}_f , \mathbf{U}_s , successively top-down/bottom-up update $\{\mathbf{R}_i\}_{i=1}^V$, \mathbf{R}_f , \mathbf{R}_s
 5: **until** Q is not decreasing
 6: **D-step:** fix $\{\mathbf{R}_v\}_{v=1}^V$, \mathbf{R}_f , \mathbf{R}_s , update $\{\mathbf{U}_i\}_{i=1}^V$, \mathbf{U}_f , \mathbf{U}_s
 7: **until** Q is converged
 // Solution
 8: **return** $\{\mathbf{U}_i\}_{i=1}^V$, \mathbf{U}_f , \mathbf{U}_s , $\{\mathbf{R}_i\}_{i=1}^V$, \mathbf{R}_f , \mathbf{R}_s

where the size of \mathbf{H}_f , \mathfrak{H} , and \mathbf{H}_s are $M \times N$, $\sum_{v=1}^V M_v \times VN$ and $M_s \times N$, and \mathfrak{H} contains V diagonal blocks $\{\mathbf{H}_v\}_{v=1}^V$.

We then successively update the three label indicator matrices \mathbf{R}_f , \mathfrak{R} , and \mathbf{R}_s when fixing the cluster centroid matrices \mathbf{U}_f , $\{\mathbf{U}_v\}_{v=1}^V$, and \mathbf{U}_s . To minimize (2.5), the following label indicator matrices update criteria will be adopted, $\forall n = 1, 2, \dots, N$,

$$r_{mn}^{(f)} = \begin{cases} 1 & m = \arg \min_k h_{kn}^{(f)} \\ 0 & \text{otherwise} \end{cases}, \quad (2.14)$$

$$r_{mn}^{(v)} = \begin{cases} 1 & m = \arg \min_k h_{kn}^{(v)} \\ 0 & \text{otherwise} \end{cases}, \quad (2.15)$$

$$r_{mn}^{(s)} = \begin{cases} 1 & m = \arg \min_k h_{kn}^{(s)} \\ 0 & \text{otherwise} \end{cases}, \quad (2.16)$$

where $h_{kn}^{(f)}$, $r_{mn}^{(f)}$, $h_{kn}^{(v)}$, $r_{mn}^{(v)}$, $h_{kn}^{(s)}$, and $r_{mn}^{(s)}$ are the entries of \mathbf{H}_f , \mathbf{R}_f , \mathbf{H}_v , \mathbf{R}_v , \mathbf{H}_s and \mathbf{R}_s , respectively. As long as the objective function of (2.5) is decreasing, \mathbf{R}_v and \mathbf{R} can be continually refined, followed by the bottom-up updates of \mathbf{R}_f and \mathbf{R}_s .

Furthermore, provided the label indicator matrices \mathbf{R}_f , \mathbf{R} , and \mathbf{R}_s , the corresponding centroid matrices \mathbf{U}_f , $\{\mathbf{U}_v\}_{v=1}^V$, and \mathbf{U}_s can be updated, and so as the corresponding distortion matrices \mathbf{D}_f , $\{\mathbf{D}_v\}_{v=1}^V$, and \mathbf{D}_s , which will also make the objective function of (2.5) decrease.

Eventually, we propose a visual pattern discovery method with multi-context-aware clustering (MCAC) in Algorithm 1. This algorithm is convergent since the solution spaces of \mathfrak{R} , \mathbf{R}_f , and \mathbf{R}_s are discrete and finite, and the objective function (2.5) is monotonically decreasing at each step. Clearly, the proposed MCAC will be degenerated to the visual pattern discovery method with spatial context-aware

clustering (SCAC) [11] if there is only one type of features and we set $\lambda_f = 0$ in (2.5) to remove the Q_β term. The complexity of the proposed algorithm is similar to k -means clustering, since our method only needs a finite run of k -means clustering.

2.3 Experiments

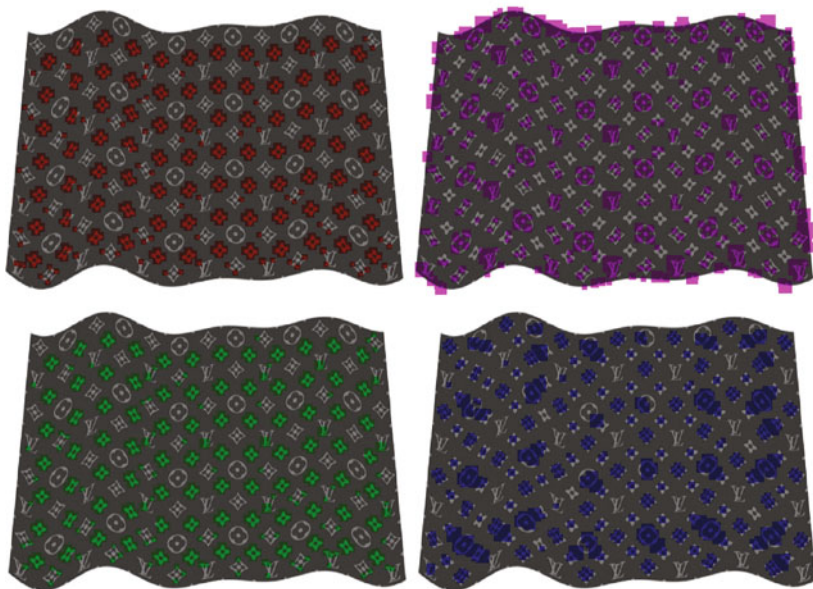
In the experiments, we set $M_v = M_f, \forall i = 1, 2, \dots, V$ for the proposed MCAC. Besides, to help parameter tuning, we let $\lambda_f = \tau_f |Q_\alpha^0 / Q_\beta^0|$ and $\lambda_s = \tau_s |Q_\alpha^0 / Q_\gamma^0|$, where Q_X^0 ($X = 1, 2, \alpha, \beta, \gamma$) is the initial value of Q_X defined by (2.5), and the nonnegative constants τ_f and τ_s are the auxiliary parameters to balance the influences from feature co-occurrences and spatial co-occurrences, respectively.

2.3.1 Spatial Visual Pattern Discovery

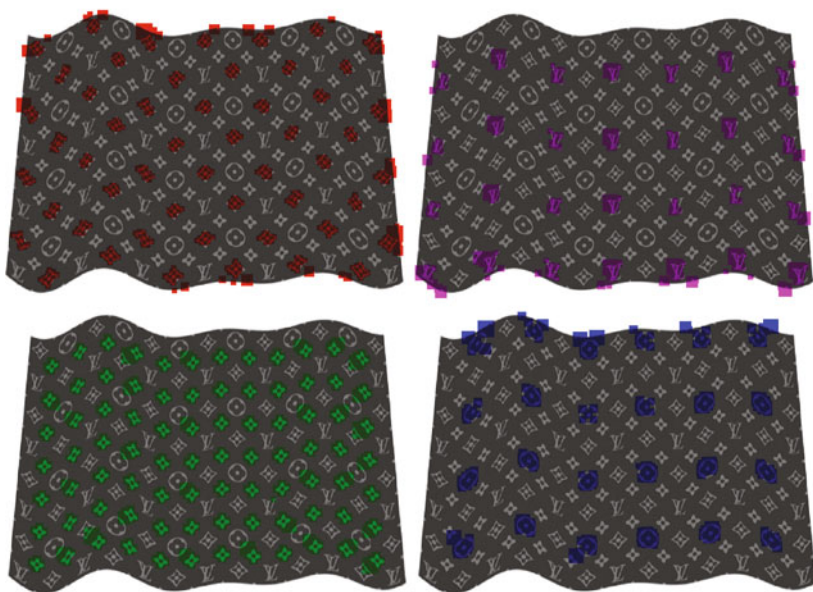
Given a single image, we detect visual primitives $\mathcal{X} = \{x_n\}_{n=1}^N$ and use one or more (e.g., V types of) features to depict each of them. Next, we apply spatial K -NN groups to build spatial context group database $\{\mathcal{G}_n^{(s)}\}_{n=1}^N$. After that, we conduct spatial pattern discovery using SCAC and the proposed MCAC. The results are shown in Figs. 2.2 and 2.3.

As shown in Fig. 2.2, the test image is a mono-colored LV monogram fabric image. Because of cloth warping, the monogram patterns are deformed, which makes pattern discovery more challenging. We detect 2604 image patches as visual primitives and use SIFT features to describe them [3]. To build spatial context groups, K -NN with $K = 8$ is applied. Other parameters are set as $M_1 = 20$, $M_s = 4$, $\tau_s = 1$ for SCAC. In Fig. 2.2, we use different colors to indicate different (4 in total) discovered spatial patterns. It is interesting to notice that SCAC can locate the monogram patterns of different spatial structures. In comparison, without considering spatial contexts of visual primitives, k -means clustering cannot obtain satisfactory results.

A comparison between SCAC and MCAC is shown in Fig. 2.3, where 422 image patches [3] are extracted. In SCAC, SIFT features [3] are used to describe these patches. While in MCAC, the patches are represented by SIFT features [3] and color histograms (CHs) [2]. Both methods construct spatial context groups by K -NN with $K = 12$ and aim to detect three categories of spatial patterns: human faces, text logos, and background edges. We highlight the instances of each discovered spatial pattern. The 1st column shows the results of SCAC, and the used parameters are as follows: $M_1 = 10$, $M_s = 3$, $\tau_s = 0.8$. The results of the 2nd column is based on MCAC, and the used parameters are as follows: $M_v = 10$, $\forall i = 1, 2$, $M_f = 10$, $M_s = 3$, $\tau_f = 1.5$, $\tau_s = 0.8$. The results show that the discovered patterns are more accurate when using MCAC. Particularly, there are more confusions between face patterns and edge patterns using SCAC than those using MCAC.



Discovered patterns using k -means clustering ($k = 4$)



Discovered patterns using SCAC.

Fig. 2.2 Pattern discovery from a mono-colored LV monogram picture. © [2014] IEEE. Reprinted, with permission, from Ref. [8]



Fig. 2.3 Pattern discovery from a colored group photograph. © [2014] IEEE. Reprinted, with permission, from Ref. [8]

2.3.2 Image Region Clustering Using Multiple Contexts

To evaluate how much feature contexts and spatial contexts can improve the clustering performance, we perform image region clustering on MSRC-V2 dataset [10]. The ground-truth labeling of MSRC-V2 is provided by [4]. As shown in Fig. 2.4, we collect five region compositions for experiments. To distinguish different region segmentations, multiple features have to be fused. Taking Fig. 2.5 as an example, while color feature can distinguish *sheep* and *cow*, it cannot distinguish *aeroplane*, *boat*, or *bicycle*. Therefore, we describe each region segmentation with the following three features: color histogram (CH), texon histogram (TH) [2], and pyramid of HOG (pHOG) [1]. The feature dimensions of CH, TH, and pHOG are 69, 400, and 680, respectively. Given an image region, all other regions in the same image are considered as in its spatial context group. Each scene category has its own region



Fig. 2.4 Sample images of five region compositions: “sheep+grass,” “cow+grass,” “aero-plane+grass+sky,” “boat+water,” and “bicycle+road”

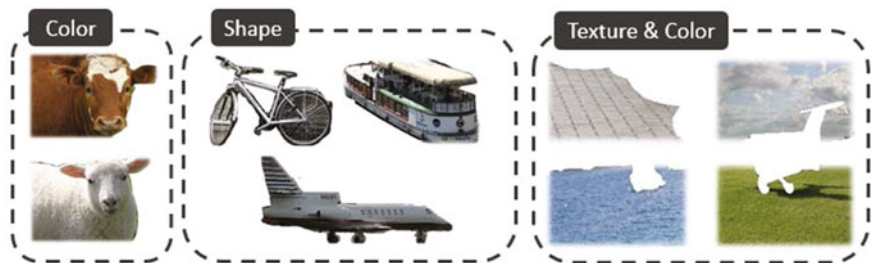


Fig. 2.5 Illustration of different features used to distinguish different region segmentations. © [2013] IEEE. Reprinted, with permission, from Ref. [9]

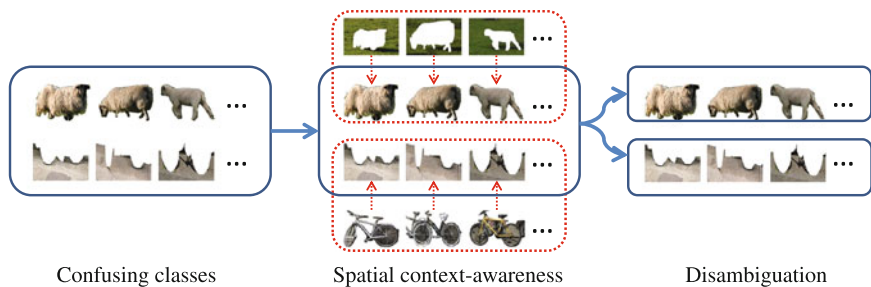


Fig. 2.6 Class disambiguation by using spatial contexts. © [2014] IEEE. Reprinted, with permission, from Ref. [8]

compositions and our goal is to cluster image regions by leveraging the spatial co-occurrence patterns. For example, visual features may suffer from the confusion between “sheep” class and “road” class as shown in Fig. 2.6, where the “sheep” regions are mislabeled as the “road” class. However, by exploring spatial contexts of image regions, the proposed MCAC are expected to better distinguish the two classes. Specifically, “grass” regions are in favor of labeling their co-occurring image regions as the “sheep” class, and similarly, the “bicycle” regions with correct labels can support the co-occurring “road” regions.

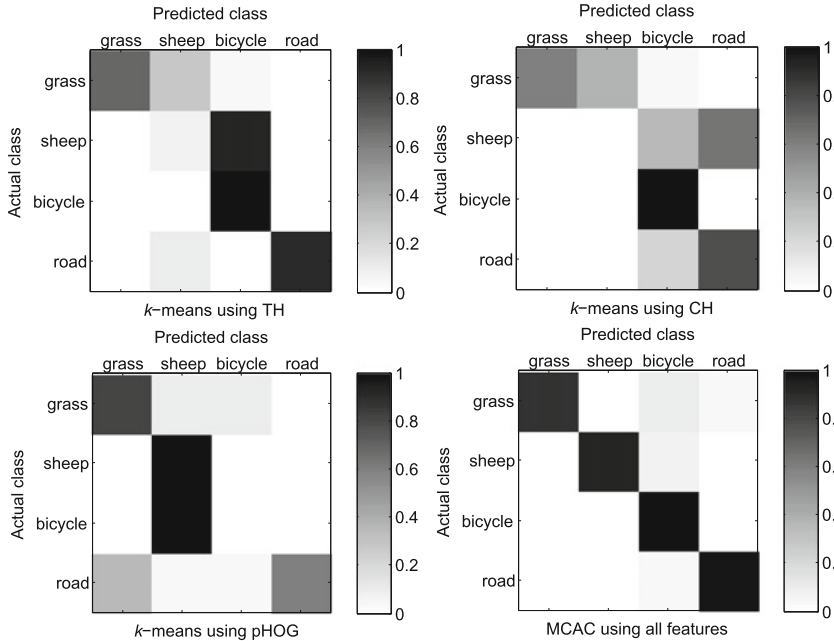


Fig. 2.7 Confusion matrices of clustering on four categories of regions. © [2014] IEEE. Reprinted, with permission, from Ref. [8]

For evaluation, we first experiment on a subset of images with two region pairs that often appear together: “sheep+grass” and “bicycle+road.” Sample images are shown in the leftmost and rightmost columns of Fig. 2.4. Each region pair has 27 image instances. There are in total 31 sheep regions, 32 grass regions, 27 bicycle regions, and 32 road regions. Because the spatial contexts of a region are the regions occurring in the same image, the spatial contextual relations only appear between regions of “sheep” and “grass” or regions of “bicycle” and “road.” We show the confusion matrices of k -means clustering and our multi-context-aware clustering in Fig. 2.7. The parameters used are as follows: $k = 4$ for k -means clustering; and $M_v = 4$, $\forall v = 1, 2, 3$, $M_f = 4$, $M_s = 2$, $\tau_f = 3.5$, $\tau_s = 1$ for MCAC. We observe that k -means clustering easily mislabeled “bicycle” as “sheep” when using TH features. This is because these TH features encode the texture of regions, and “sheep” regions have similar texture to “bicycle” regions. When using CH features, it is easy to mislabel “sheep” regions as “road” regions because of their similar colors. Also, with similar shape features, quite a lot of “sheep” regions are mislabeled as “bicycle” class when using pHOG features. Besides the limited description ability of a single type of feature, as k -means does not consider the spatial dependencies among regions, it also causes confusions among different classes. By considering the feature co-occurrences of CH, TH and pHOG, and the spatial co-occurrences of “sheep” and “grass” regions, as well as “bicycle” and “road” regions, the proposed MCAC can well improve the

Table 2.1 Results of image region clustering on the MSRC-V2 subset, sample images of which are shown in Fig. 2.4. Based on Ref. [8]

Method	Error(%)
k -means clustering using TH	44.31
k -means clustering using CH	55.21
k -means clustering using pHOG	47.63
k -means clustering using TH+CH+pHOG	38.39
MCAC using all features	29.86

clustering results on individual features and finally reduce the confusion among the region classes. Specifically, our method can leverage the “grass” regions to correct the confused “sheep” regions and vice versa. A similar improvement can be observed for “bicycle” and “road.”

In the above experiment, we show the advantage of the proposed MCAC in dealing with image regions of clear spatial contexts. However, Fig. 2.4 shows image regions may have ambiguous spatial contexts, which will also be used to evaluate the proposed method. Specifically, we collect 30 “sheep+grass,” 29 “cow+grass,” 30 “aeroplane+grass+sky,” 31 “boat+water,” and 30 “bicycle+road.” The numbers of “sheep,” “grass,” “cow,” “sky,” “aeroplane,” “boat,” “water,” “bicycle,” and “road” are 34, 104, 34, 53, 30, 47, 39, 30, and 51, respectively. Notice that in this challenging dataset, different image regions may share the same spatial context. For example, “grass” occurs in three different scenes: “sheep+grass,” “cow+grass,” and “aeroplane+grass+sky.”

The results of k -means clustering and MCAC are shown in Table 2.1, where the same 10% seeds per category from ground truth are randomly chosen for initialization. The clustering error rate of the proposed MCAC is 29.86%. It brings a considerable improvement than the best one (i.e., 33.65%) obtained by k -means clustering on the individual features or the concatenated multiple features. We can obtain similar observation in terms of average of precision and average of recall. In k -means clustering, we set $k = 9$ as there are 9 different types of image regions. Parameters used in MCAC are $M_v = 9$, $\forall i = 1, 2, 3$, $M_f = 9$, $M_s = 5$, $\tau_f = 3.5$, $\tau_s = 1$.

Some representative clustering results of the proposed MCAC are shown in Fig. 2.8. Despite large intra-class variations, our method can still obtain a satisfactory clustering result by using both spatial and feature contexts. For example, the “cow” regions are with different colors and perspectives. We also note that there may contain “water” regions in some “sheep+grass” and “cow+grass” region compositions. These small amounts of “water” regions are mislabeled as “grass” class because of its preference of “cow”/“sheep” contexts. Moreover, because the feature appearance and spatial contexts are similar, there still exist confusions between a few regions of “sheep” and “cow,” “bicycle” and “sheep,” “boat” and “aeroplane,” “water” and “sky,” “boat” and “bicycle,” and “water” and “road.” Nevertheless, the mislabeling results are only among the minority.







































Category	True positive			False positive
Grass				
Cow				
Sheep				 
Sky				
Aeroplane				
Water				
Bicycle				 
Road				
Boat				

Fig. 2.8 Exemplar clustering results of MCAC. © [2014] IEEE. Reprinted, with permission, from Ref. [8]

2.4 Summary of this Chapter

Because of the structure and content variations of complex visual patterns, they greatly challenge most existing 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 dependencies among visual primitives and simply performing k -means clustering to obtain the visual vocabulary, we explore 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. Our proposed self-learning procedure can iteratively refine the pattern discovery results and guarantee to converge. Furthermore, we explore feature contexts and utilize the co-occurrence patterns among multiple types of features to handle the content variations of visual patterns. By

doing so, our method can leverage multiple types of features to further improve the performance of clustering and pattern discovery. The experiments on spatial visual pattern discovery and image region clustering validate the advantages of the proposed method.

References

1. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the International Conference on Image and Video Retrieval, pp. 401–408 (2007)
2. Lee, Y., Grauman, K.: Object-graphs for context-aware visual category discovery. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(2), 346–358 (2012)
3. Lowe, D.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
4. Malisiewicz, T., Efros, A.: Improving apatial support for objects via multiple segmentations. In: Proceedings of British Machine Vision Conference, vol. 2 (2007)
5. Russell, B., Freeman, W., Efros, A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1605–1614 (2006)
6. Su, Y., Jurie, F.: Visual word disambiguation by semantic contexts. In: Proceedings of IEEE International Conference on Computer Vision, pp. 311–318 (2011)
7. Tuytelaars, T., Lampert, C., Blaschko, M., Buntine, W.: Unsupervised object discovery: a comparison. *Int. J. Comput. Vis.* **88**(2), 284–302 (2010)
8. Wang, H., Yuan, J., Wu, Y.: Context-aware discovery of visual co-occurrence patterns. *IEEE Trans. Image Process.* **23**(4), 1805–1819 (2014)
9. Weng, C., Wang, H., Yuan, J.: Hierarchical sparse coding based on spatial pooling and multi-feature fusion. In: Proceedings of the IEEE International Conference on Multimedia Expo, pp. 1–6 (2013)
10. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Proceedings of IEEE International Conference on Computer Vision (2005)
11. Yuan, J., Wu, Y.: Context-aware clustering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
12. Yuan, J., Wu, Y.: Mining visual collocation patterns via self-supervised subspace learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **42**(2), 1–13 (2012)
13. Yuan, J., Wu, Y., Yang, M.: From frequent itemsets to semantically meaningful visual patterns. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 864–873 (2007)

Visual Pattern Discovery and Recognition

Wang, H.; Weng, C.; Yuan, J.

2017, X, 87 p. 33 illus., 9 illus. in color., Softcover

ISBN: 978-981-10-4839-5