Context-Aware Discovery of Visual Co-Occurrence Patterns

Hongxing Wang, *Student Member, IEEE*, Junsong Yuan, *Member, IEEE*, and Ying Wu, *Senior Member, IEEE*

***Abstract*— Once an image is decomposed into a number of visual primitives, e.g., local interest points or regions, it is of great interests to discover meaningful visual patterns from them. Conventional clustering of visual primitives, however, usually ignores the spatial and feature structure among them, thus cannot discover high-level visual patterns of complex structure. To overcome this 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 better address the ambiguities of clustering visual primitives. We formulate the pattern discovery problem as a regularized *k*-means clustering where spatial and feature contexts are served as constraints to improve the pattern discovery results. A novel self-learning procedure is proposed to utilize the discovered spatial or feature patterns to gradually refine the clustering result. Our self-learning procedure is guaranteed to converge and experiments on real images validate the effectiveness of our method.**

***Index Terms*— Clustering, feature context, spatial context, visual pattern discovery.**

I. Introduction

IMAGES can be decomposed into visual primitives, *e.g.*, local interest points or image regions. With each visual primitive 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. Although it has been a common practice to build a visual vocabulary for image analysis, it remains a challenging problem to cluster visual primitives into meaningful patterns. Most existing visual primitive clustering methods ignore the spatial structure among the visual primitives [1], thus bring unsatisfactory results. 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 [2], [3].

As visual primitives are not independent of 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 [4], [5], *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. On the other hand, even though two visual primitives share similar features, they may not belong to the same visual pattern if their spatial contexts are completely different. By considering the spatial dependency among visual primitives, many previous work have tried to discover the spatial co-occurrence patterns of visual primitives, *e.g.*, visual phrases [6]–[8]. Such co-occurrence patterns, once discovered, can be utilized to better address the ambiguities and uncertainties among visual primitives and have proven to be helpful for object categorization and image search.

Despite previous success of utilizing spatial visual patterns, there lacks a principled solution that can leverage visual patterns to improve the clustering of visual primitives. For example, [6] applies data mining methods first to find the spatial visual patterns. After that it relies on a separate subspace learning to refine the presentation and clustering of visual primitives. As the clustering of visual primitives and the discovery of visual pattern will influence each other, it is preferred to provide a uniformed solution that can integrate visual pattern discovery into the process of visual primitives clustering. Moreover, besides the spatial dependencies among visual primitives, a visual pattern can exhibit certain feature dependencies as well. For example, instead of using a single feature vector, a sheep image region (as a visual primitive) can also be described by a co-occurrence of the white (color) and fur (texture) attributes. Since visual primitives can be easily described by multiple types of features or attributes, besides discovering spatial co-occurrence patterns, it is equally interesting to discover feature/attribute co-occurrence patterns in an image. Such a feature/attribute co-occurrence discovery problem, however, is seldom addressed in the literature.

In this work, we propose a principled solution to address the above mentioned two problems of visual co-occurrence pattern discovery. First, we propose a regularized *k*-means formulation that adds spatial co-occurrences as the constraints to the conventional *k*-means clustering. It can leverage the discovered spatial co-occurrence patterns to correct the labeling of visual primitives. A novel self-supervised clustering procedure is proposed to allow the co-occurrence pattern discovery and visual primitive clustering to help each other, thus leads to a better visual vocabulary as well as better spatial

Manuscript received March 6, 2013; revised November 29, 2013; accepted February 2, 2014. Date of publication February 25, 2014; date of current version March 7, 2014. This work was supported by the Nanyang Assistant Professorship SUG under Grant M4080134. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hassan Foroosh.

H. Wang and J. Yuan are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: hwang8@e.ntu.edu.sg; jsyuan@ntu.edu.sg).

Y. Wu is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: yingwu@northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2308416

Algorithm 1 Spatial Context-Aware Clustering

```
input  : database $\mathcal{D} = \{v_i\}_{i=1}^{N}$;
         spatial contextual relations $\mathbf{Q}_\mathrm{s}$;
         parameters: $M$, $M_\mathrm{s}$, $\lambda_\mathrm{s}$
output : feature word and spatial pattern lexicons: $\mathbf{\Omega}$ ($\mathbf{U}$) and
         $\mathbf{\Psi}_\mathrm{s}$ ($\mathbf{U}_\mathrm{s}$); clustering results $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$
Init: (1) cluster data samples to obtain $\mathbf{U}$;
      (2) based on $\mathbf{U}$, cluster spatial context groups to obtain $\mathbf{U}_\mathrm{s}$;
while not converged do
    E-step: fix $\mathbf{U}$ and $\mathbf{U}_\mathrm{s}$, update $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$
        nested-E step: fix $\mathbf{R}$, update $\mathbf{R}_\mathrm{s}$ (Eq. 11)
        nested-M step: fix $\mathbf{R}_\mathrm{s}$, update $\mathbf{R}$ (Eq. 13)
    if J is decreasing then
        goto E-setp
    else
        goto M-step
    M-step: fix $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$, update $\mathbf{U}$ and $\mathbf{U}_\mathrm{s}$ separately
return $\mathbf{U}$, $\mathbf{U}_s$, $\mathbf{R}$, $\mathbf{R}_s$.
```

co-occurrence patterns. Moreover, besides considering the spatial co-occurrence pattern, we also add the feature co-occurrence pattern into consideration and provide a uniformed formulation that can handle both spatial and feature co-occurrence patterns. Both spatial and feature structure of the visual patterns can thus be exhibited. To the best of our knowledge, it is the first work that can consider both spatial co-occurrences and feature co-occurrences with a uniformed formulation.

The contributions of our work are summarized in the following two aspects.

- We provide a regularized *k*-means clustering formulation of visual co-occurrence pattern discovery, and propose a novel self-supervised learning procedure to leverage the discovered visual co-occurrence patterns to guide the clustering of visual primitives. Such a self-supervised learning is proven to converge in finite steps.
- We further extend the spatial co-occurrence pattern discovery to feature co-occurrence pattern discovery. A uniformed formulation is proposed that can leverage both spatial and feature co-occurrence patterns and let them boost each other to improve the pattern discovery results.

The rest of this paper is organized as follows: Sec. II discusses the related work of the paper. In Sec. III, we present the formulation of the spatial context-aware clustering and apply the proposed Algorithm 1 for an effective data simulation. In Sec. IV, we propose Algorithm 2 to combine feature contexts and spatial contexts for clustering improvement and pattern discovery. More experiments are presented in Sec. V. Finally, we conclude the paper in Sec. VI. This manuscript expands upon our previous conference papers [9] and [10].

II. Related Work

An information-rich visual pattern is composed of a group of visual primitives with a certain spatial configuration. The discovery of visual patterns can contribute to many applications such as image search [8], [11]–[13], object categorization [6], [7], [14]–[16], scene understanding [17]–[25], and video analysis [26]–[29]. Many approaches have been proposed to discover such frequent spatial patterns of visual primitives. These methods can be generally divided into bottom-up visual pattern mining and top-down generative visual pattern modeling.

For the bottom-up visual pattern mining methods, the spatial co-occurrence is an underlying principle. The early studies usually rely on the transaction-based neighborhood representation, which encodes which visual words appear together in a given spatial neighborhood. For example, the representative methods include [6], [27], [30]. The method proposed in [27] performs clustering for visual pattern discovery. The frequent pattern mining methods [31] are used in [6] and [30].

Since the frequency information of visual words is usually not included in the transaction representation, several methods propose to consider the word frequencies together with spatial neighborhood representation such as in [32] and [33]. In [32], the authors propose a bag-to-set (B2S) representation to transform word frequencies into a longer transaction. However this representation cannot avoid the generation of artificial visual patterns. Recently the bag of frequent local histograms method proposed in [33] keeps all the frequency information without bringing artificial patterns.

The above methods are all proposed to mine frequent visual word compositions. However, we have to first quantize visual features of visual primitives to obtain the visual word vocabulary. A poor word quantization may degrade the visual pattern mining performance. Therefore there are also some word-free visual pattern mining methods such as in [21], [34], and [35]. In [21], each image is randomly partitioned several times to generate a pool of subimages, followed by the common visual pattern discovery from the frequent appearing subimages. In [34], by feature indexing and matching, a hierarchical representation of spatially feature compositions are learned. The method proposed in [35] utilizes the property that the instances from the same visual pattern can be generated by each other to develop a visual category pattern discovery method called clustering by composition.

Besides the spatial co-occurrence visual pattern discovery, some researchers focus on discovering geometry preserving visual patterns. In [36], the authors use the relative spatial distance, relative scale difference and relative headings between pairwise visual words to represent the geometric relationship between them. The consistent geometric visual patterns are discover by a further frequent subgraph pattern mining. In [8], the authors derive an offset space by the relative location difference between each visual word pair in two images. Because the visual words exhibiting the same geometric relationship will be mapped to the same place on the offset space, the high-order geometry preserving visual patterns can be mined by Hough Voting.

In addition to the above bottom-up visual pattern mining, there are also considerable methods in modeling visual patterns from top down, such as the constellation model in [37], the pictorial model in [38], the bayesian model in [39], the tree model in [40] and [41], the active basis model in [42], the hierarchical model in [43]–[45], the conditional random field model in [46]. Specially, in [37], object classes are represented as constellations of visual parts in a probabilistic model. In [38], the visual parts of a category of objects are modeled with deformable configuration in a pictorial structure. In [39], the spatial compositional structure of visual objects is learned using Bayesian network. In [40], object categories are detected by similar geometric, photometric and topological properties in a tree model. Another tree-structured graphical model proposed in [41] learns the object co-occurrences and spatial dependency relationships. In [42], it learns frequent occurring spatial structures in natural images using active basis model. The hierarchical spatial patterns presented in images are studied in [43]–[45]. The configurations of objects in scenes are discovered by using a conditional random field in [46].

In the previous approaches, visual patterns are discovered either via top down generative model, or from bottom up mining. However, the composition of visual patterns and the clustering of visual primitives influence each other. Therefore, a principled solution that can integrate the top-down and bottom-up processes is preferred [4], [5], where it can leverage visual patterns to improve the clustering of visual primitives. Moreover, besides spatial composition of visual primitives, because of the feature dependencies of a visual pattern across multiple types of features, there also needs to utilize feature co-occurrences of visual primitives for an enhanced discovery of visual patterns and clustering of visual primitives. But much less work can address such feature co-occurrence pattern mining. To address the problems mentioned above, we propose a self-supervised clustering procedure that can allow the spatial co-occurrence pattern discovery, feature co-occurrence discovery and visual primitive clustering to help one another.

III. Spatial Context-Aware Clustering

A. Motivating Example: Clustering Visual Primitives

We illustrate our spatial context-aware clustering in a case study of clustering visual primitives. Each visual primitive is described by a feature vector $v := \mathbf{f}$, located at (x, y) in the image space. In general $\mathbf{f} \in \mathbb{R}^d$ can be any visual features to characterize a local image region, like color histograms or SIFT-like features [47]. An image is a collection of visual primitives, and we denote the visual primitive database as $\mathcal{D}_v = \{v_n\}_{n=1}^N$. After clustering these visual primitives into feature words, we can label each $v_n \in \mathcal{D}_v$ with $l(v_n) \in \mathbf{\Omega}$, where $\mathbf{\Omega}$ is the feature word lexicon of size $|\mathbf{\Omega}| = M$.

The spatial contexts of a visual primitive are its local spatial neighbors in the image, *i.e.*, those visual primitives that collocate with it. Take human face image for example, the spatial contexts of a nose contain two eyes and a mouth. For each visual primitive $v_n \in \mathcal{D}_v$, we define its local spatial neighborhood, *i.e.*, K-nearest neighbors (K-NN) or ϵ-nearest neighbors (ϵ-NN), as its *spatial context group* $\mathcal{G}_n^{(s)} = \{v_n, v_{n_1}, v_{n_2}, \ldots, v_{n_K}\}$. The *spatial context database* is denoted by $\mathbf{G}_s = \{\mathcal{G}_n^{(s)}\}_{n=1}^N$. Once the visual primitives are labeled by $\mathbf{\Omega}$, the spatial context group database $\mathbf{G}_s$ can be transferred to a *spatial context transaction database* with N records. And we have the definition as follows.

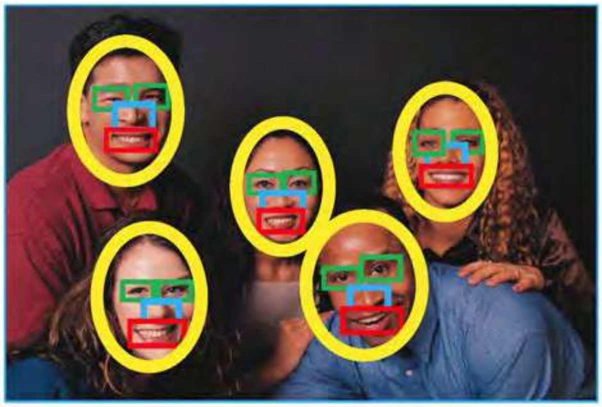


Fig. 1. A spatial pattern *{left eye, right eye, nose, mouth}* co-occur frequently which forms multiple faces of different people.

Definition 1 (Spatial context transaction). *The spatial context transaction of the visual primitive v_n refers to the co-occurrences of different categories of visual primitives appearing in the spatial context group of v_n.*

The spatial context transaction database can also be represented as a sparse integer matrix $\mathbf{T}_s \in \mathbb{R}^{M \times N}$, where each column is a spatial context transaction $\mathbf{t}_n^{(s)} \in \mathbb{Z}^M$. The entry $t_{mn}^{(s)} = x$ indicates that the n^{th} transaction contains x visual primitives belonging to the m^{th} feature word. In the case of using spatial K-NN to define context group, we have $\sum_{m=1}^M t_{mn}^{(s)} = K, \ \forall\, n = 1, \ldots, N$, because each context group $\mathcal{G}_n^{(s)}$ contains K visual primitives.

A sparse binary matrix $\mathbf{Q}_s \in \mathbb{R}^{N \times N}$ can be used to describe the spatial context relations among the visual primitives, where $q_{ij}^{(s)} = 1$ denotes that v_i belongs to the context group of v_j, *i.e.*, $v_i \in \mathcal{G}_j^{(s)}$; and $q_{ij}^{(s)} = 0$ otherwise. The matrix $\mathbf{Q}_s$ is symmetric when using ϵ-NN to define spatial neighbors, while an asymmetric matrix when using K-NN. The context matrix $\mathbf{Q}_s$ plays a critical role in our spatial context-aware clustering as it introduces extra relations among visual primitives other than those in the feature space.

Based on the feature word lexicon $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = M$), we can further define a *spatial pattern lexicon* $\Psi_s = \{\mathcal{P}_m^{(s)}\}_{m=1}^{M_s}$. A spatial pattern is composed of a collection of feature words. This implies $\mathcal{P}_m^{(s)} \subset \mathbf{\Omega}$. Compared with feature words which label visual primitives $\mathcal{D}_v$, spatial patterns label spatial context transactions $\mathbf{T}_s$, *i.e.*, label spatial context groups $\mathbf{G}_s$. As spatial pattern describes the spatial dependencies among feature words, they can present more meaningful patterns in a higher level [6]. For example in Fig. 1, the existence of a spatial pattern $\mathcal{P} = \{left\ eye, right\ eye, nose, mouth\}$ shows that these four categories of visual primitives: left eye, right eye, nose and mouth in $\mathbf{\Omega}$ co-occur frequently and form a meaningful visual pattern, *i.e.*, face. Moreover, the discovered visual patterns can refine the primitives clustering of uncertainty, which will also be discussed in detail in Sec. III-B and Sec. III-C. We represent $\mathcal{P}_m^{(s)} \in \Psi_s$ as an integer vector $\mathbf{u}_m^{(s)} \in \mathbb{Z}^M$ which describes its word compositions, where $\mathbf{u}_m^{(s)}(i) = x$ indicates that the i^{th} word is contained in $\mathcal{P}_m^{(s)}$ and occurs x times. The matrix $\mathbf{U}_s \in \mathbb{R}^{M \times M_s}$

is further applied to represent Ψ_{s}, where each column of $\mathbf{U}_s$ is a $\mathbf{u}_m^{(\text{s})}$. Correspondingly, we use a real matrix $\mathbf{U} \in \mathbb{R}^{d \times M}$ to represent $\mathbf{\Omega}$, where each column is a real vector to represent a feature word prototype $\mathbf{u}_m \in \mathbb{R}^d$.

Above, we describe visual pattern discovery in a single image. However it can be easily extended to multiple images. For example, we can treat multiple images as one huge image and do not allow visual primitives from different images to compose visual patterns.

B. Problem Formulation

We first revisit the k-means clustering of visual primitives $\mathcal{D}_v = \{v_n\}_{n=1}^N$, where the following mean square distortion needs to be minimized:

$$\mathbf{J}_1 = \sum_{m=1}^{M}\sum_{n=1}^{N} r_{mn}\|\mathbf{u}_m - \mathbf{f}_n\|^2 = tr(\mathbf{R}^{\text{T}}\mathbf{D}), \tag{1}$$

where

- $\mathbf{f}_n$ is the $d \times 1$ feature vector; $\mathbf{u}_m$ is the center of the cluster (prototype of feature words); $\|\cdot\|$ denotes the Euclidean distance and $tr(\cdot)$ denotes the matrix trace;
- $\mathbf{D}_{M\times N}$ denotes the distance matrix, where each element $d_{mn} = \|\mathbf{u}_m - \mathbf{f}_n\|^2$ denotes the distance between the n^{th} visual primitives and the m^{th} feature word prototype;
- $\mathbf{R}_{M\times N}$ denotes the label indicator matrix of the visual primitives, where $r_{mn} = 1$ if the n^{th} visual primitive is labeled with the m^{th} word; and $r_{mn} = 0$ otherwise.

Standard EM-algorithm can be performed to minimize the distortion in Eq. 1 by iteratively updating $\mathbf{R}$ (E-step) and $\mathbf{D}$ (M-step). However, by minimizing the objective function $\mathbf{J}_1$, k-means clustering tries to maximize the data likelihood under mixture Gaussian distribution and assumes all observation samples $v_n \in \mathcal{D}_v$ are independent from one another in the feature space:

$$Pr(\mathcal{D}_v|\mathbf{\Omega}) = \prod_{n=1}^{N} Pr(v_n|\mathbf{\Omega}). \tag{2}$$

Such an independent assumption, however, does not hold here because visual primitives have spatial dependencies with each other.

In order to consider spatial dependency for clustering, we propose a regularized k-means to minimize:

$$\begin{aligned}\mathbf{J} &= \sum_{m=1}^{M}\sum_{n=1}^{N} r_{mn}\|\mathbf{u}_m - \mathbf{f}_n\|^2 \\ &\quad + \lambda_s \sum_{m=1}^{M_{\text{s}}}\sum_{n=1}^{N} r_{mn}^{(\text{s})} d_H(\mathbf{u}_m^{(\text{s})}, \mathbf{t}_n^{(\text{s})}) \\ &= tr(\mathbf{R}^{\text{T}}\mathbf{D}) + \lambda_{\text{s}} tr(\mathbf{R}_{\text{s}}^{\text{T}}\mathbf{D}_{\text{s}}),\end{aligned} \tag{3}$$

where

- $\lambda_{\text{s}} > 0$ is a positive constant for regularization;
- $r_{mn}^{(\text{s})}$ is the binary label indicator of transactions, with $r_{mn}^{(\text{s})} = 1$ denoting that the n^{th} spatial context transaction is labeled with the m^{th} spatial pattern; and $r_{mn}^{(\text{s})} = 0$ otherwise. Similar to $\mathbf{R}$, $\mathbf{R}_{\text{s}}$ is an $M_{\text{s}} \times N$ matrix to describe the clustering results of spatial context transactions $\{\mathbf{t}_n^{(\text{s})}\}_{n=1}^N$. For deterministic clustering, we have the following constraints for $\mathbf{R}$ and $\mathbf{R}_{\text{s}}$:

$$\sum_{m=1}^{M} r_{mn} = 1, \quad \sum_{m=1}^{M_{\text{s}}} r_{mn}^{(\text{s})} = 1, \quad \forall\, n = 1, \ldots, N. \tag{4}$$

- $d_H(\mathbf{u}_m^{(\text{s})}, \mathbf{t}_n^{(\text{s})})$ denotes the Hamming distance between two binary vectors[1]: a transaction $\mathbf{t}_n^{(\text{s})}$ and a context pattern $\mathbf{u}_m^{(\text{s})}$, where $\mathbf{1}$ is the $M \times 1$ all 1 vector:

$$\begin{aligned}&d_H(\mathbf{u}_m^{(s)}, \mathbf{t}_n^{(s)}) \\ &\quad = M - \left[(\mathbf{t}_n^{(s)})^{\text{T}}\mathbf{u}_m^{(s)} + (\mathbf{1} - \mathbf{t}_n^{(s)})^{\text{T}}(\mathbf{1} - \mathbf{u}_m^{(s)})\right] \\ &\quad = (\mathbf{t}_n^{(s)})^{\text{T}}\mathbf{1} + (\mathbf{u}_m^{(s)})^{\text{T}}\mathbf{1} - 2(\mathbf{t}_n^{(s)})^{\text{T}}\mathbf{u}_m^{(s)}.\end{aligned} \tag{5}$$

Given the objective function in Eq. 3 with M, M_{s} and λ are fixed parameters, our goals are: (1) clustering all the visual primitives v_i into M classes (feature word lexicon $\mathbf{\Omega}$) and (2) clustering all the context transactions $\mathbf{t}_n^{(\text{s})} \in \mathbf{T}_{\text{s}}$ into M_{s} classes (spatial pattern lexicon Ψ_{s}). The clustering results are presented by $\mathbf{R}$ and $\mathbf{R}_{\text{s}}$ respectively. Since each visual primitive can generate a spatial context group, we finally end up with two labels for every primitive: (1) the word label of itself and (2) the pattern label of the spatial group it generates. Compared with k-means clustering which assumes convex shape for each cluster in the feature space, *e.g.*, a mixture of Gaussian, our regularization term can modify the cluster into an arbitrary shape by considering the influences from the visual pattern level. Similar to the k-means clustering, this formulation is also a mixed integer problem with multiplicative terms where we cannot estimate $\mathbf{D}$, $\mathbf{D}_{\text{s}}$, $\mathbf{R}$ and $\mathbf{R}_{\text{s}}$ simultaneously.

C. Algorithm

The objective function in Eq. 3 includes two parts:

$$\mathbf{J} = \underbrace{tr(\mathbf{R}^{\text{T}}\mathbf{D})}_{\mathbf{J}_1} + \underbrace{\lambda_s tr(\mathbf{R}_{\text{s}}^{\text{T}}\mathbf{D}_{\text{s}})}_{\mathbf{J}_2},$$

where $\mathbf{J}_1 = tr(\mathbf{R}^{\text{T}}\mathbf{D})$ and $\mathbf{J}_2 = \lambda_s tr(\mathbf{R}_{\text{s}}^{\text{T}}\mathbf{D}_{\text{s}})$ correspond to the quantization distortions of visual primitives and spatial context groups respectively. Although it looks we could minimize $\mathbf{J}$ by minimizing $\mathbf{J}_1$ and $\mathbf{J}_2$ separately, *e.g.*, through two independent EM-processes, this is actually infeasible because $\mathbf{J}_1$ and $\mathbf{J}_2$ are coupled. By further analyzing $\mathbf{J}_1$ and $\mathbf{J}_2$, we find that although visual primitive distortions $\mathbf{D}$ only depend on $\mathbf{R}$, the spatial context group distortions $\mathbf{D}_{\text{s}}$ depend on *both* visual primitive labels $\mathbf{R}$ and spatial context group labels $\mathbf{R}_{\text{s}}$. Thus it is infeasible to minimize $\mathbf{J}_1$ and $\mathbf{J}_2$ separately due to their correlation. In the following, we show how to decouple the dependencies between $\mathbf{J}_1$ and $\mathbf{J}_2$ and propose our spatial context-aware clustering: a nested-EM algorithm.

[1] Strictly, $\mathbf{t}_n^{(\text{s})}$ and $\mathbf{u}_m^{(\text{s})}$ are binary vectors only if $\mathbf{t}_n^{(\text{s})}$ contains distinguishable primitives, *i.e.*, each primitive belongs to a different word in $\mathbf{t}_n^{(\text{s})}$. However, our solution is generic and do not need $\mathbf{t}_n^{(\text{s})}$ to be binary, as long as we apply the distortion measure (called Hamming distortion) derived from Hamming distance as in Eq. 5.

Initialization:

1) Clustering all visual primitives $\{v_n\}_{n=1}^N$ into M classes, *e.g.*, through k-means clustering, based on the Euclidean distance.
2) Obtaining the feature word lexicon $\boldsymbol{\Omega}$ (represented by $\mathbf{U}$) and the distortion matrix $\mathbf{D}$.
3) Clustering all spatial context groups $\{\mathcal{G}_n^{(\mathrm{s})}\}_{n=1}^N$ into M_s classes based on the Hamming distance, and obtaining the spatial pattern lexicon Ψ_s (represented by $\mathbf{U}_\mathrm{s}$), as well as the distortion matrix $\mathbf{D}_\mathrm{s}$.

E-step:

The task is to label visual primitives $\{v_n\}_{n=1}^N$ with $\boldsymbol{\Omega}$ and spatial context groups $\{\mathcal{G}_n^{(\mathrm{s})}\}_{n=1}^N$ with Ψ, namely to update $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$ given $\mathbf{D}$ and $\mathbf{D}_\mathrm{s}$, where $\mathbf{D}$ and $\mathbf{D}_\mathrm{s}$ can be directly computed from $\mathbf{U}$ and $\mathbf{U}_\mathrm{s}$, respectively. Based on the analysis above, we need to optimize $\mathbf{R}$ (corresponding to $\mathbf{J}_1$) and $\mathbf{R}_\mathrm{s}$ (corresponding to $\mathbf{J}_2$) *simultaneously* to minimize $\mathbf{J}$, because $\mathbf{J}_1$ and $\mathbf{J}_2$ are correlated.

According to the Hamming distortion in Eq. 5, we can derive the matrix form of spatial context group distortions:

$$\mathbf{D}_\mathrm{s} = -2\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{T}_\mathrm{s} + \mathbf{1}_{\mathbf{T}_\mathrm{s}}\mathbf{T}_\mathrm{s} + \mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}}, \tag{6}$$

where $\mathbf{1}_{\mathbf{T}_\mathrm{s}}$ is an $M_\mathrm{s} \times M$ all 1 matrix and $\mathbf{1}_{\mathbf{U}_\mathrm{s}}$ is an $M \times N$ all 1 matrix. Moreover, transaction database $\mathbf{T}_\mathrm{s}$ can be determined by

$$\mathbf{T}_\mathrm{s} = \mathbf{R}\mathbf{Q}_\mathrm{s}. \tag{7}$$

Because each transaction column can be obtained as

$$\mathbf{t}_n^{(\mathrm{s})} = \sum_{i=1}^{N} q_{in}^{(\mathrm{s})}\mathbf{r}_i,$$

where $\mathbf{r}_i$ denotes the i^{th} column of $\mathbf{R}$ which describes the word label of v_i. Now, we derive Eq. 3 as follows:

$$\begin{aligned}\mathbf{J} &= tr(\mathbf{R}^\mathrm{T}\mathbf{D}) + \lambda_\mathrm{s} tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{D}_\mathrm{s})\\ &= tr\{\mathbf{R}_\mathrm{s}^\mathrm{T}[\lambda_\mathrm{s}(-2\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{R}\mathbf{Q}_\mathrm{s} + \mathbf{1}_{\mathbf{T}_\mathrm{s}}\mathbf{R}\mathbf{Q}_\mathrm{s} + \mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}})]\}\\ &\quad + tr(\mathbf{R}^\mathrm{T}\mathbf{D}) \qquad (8)\\ &= tr[\mathbf{R}^\mathrm{T}(\mathbf{D} - \lambda_\mathrm{s}(2\mathbf{U}_\mathrm{s}^\mathrm{T} - \mathbf{1}_{\mathbf{T}_\mathrm{s}})^\mathrm{T}\mathbf{R}_\mathrm{s}\mathbf{Q}_\mathrm{s}^\mathrm{T})]\\ &\quad + \lambda_\mathrm{s} tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}}). \qquad (9)\end{aligned}$$

Based on the above analysis, we propose an E-step to iteratively update $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$ to decrease $\mathbf{J}$. Recall that $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$ are label indicator matrices constrained by Eq. 4.

1) **Bottom-up co-occurrence pattern discovery:** We first fix visual word labeling $\mathbf{R}$ to update visual pattern labeling $\mathbf{R}_\mathrm{s}$. Based on Eq. 8, let

$$\mathbf{H}_\mathrm{s} \overset{\triangle}{=} \lambda_\mathrm{s}(-2\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{R}\mathbf{Q}_\mathrm{s} + \mathbf{1}_{\mathbf{T}_\mathrm{s}}\mathbf{R}\mathbf{Q}_\mathrm{s} + \mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}}),$$

we have

$$\mathbf{J} = \underbrace{tr(\mathbf{R}^\mathrm{T}\mathbf{D})}_{\mathbf{J}_1} + \underbrace{tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{H}_\mathrm{s})}_{\mathbf{J}_2}. \tag{10}$$

Therefore we only need to minimize $\mathbf{J}_2 = tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{H}_\mathrm{s})$ as $\mathbf{J}_1 = tr(\mathbf{R}^\mathrm{T}\mathbf{D})$ is a constant given $\mathbf{R}$ and $\mathbf{U}$. Because each column of $\mathbf{R}_\mathrm{s}$ contains a single 1 (Eq. 4), we update $\mathbf{R}_\mathrm{s}$ to minimize $\mathbf{J}_2$ based on the following criterion, $\forall\ n = 1, 2, \ldots N$:

$$r_{mn}^{(\mathrm{s})} = \begin{cases} 1 & m = \arg\min_k h_{kn}^{(\mathrm{s})} \\ 0 & otherwise, \end{cases} \tag{11}$$

where $h_{kn}^{(\mathrm{s})}$ is the element of $\mathbf{H}_\mathrm{s}$ and $r_{mn}^{(\mathrm{s})}$ is the element of $\mathbf{R}_s$. $\mathbf{H}_\mathrm{s}$ can be calculated based on $\mathbf{Q}_\mathrm{s}$, $\mathbf{U}_\mathrm{s}$ and $\mathbf{R}$ which are all given.

2) **Top-down refinement:** Similar to the above step, now we fix $\mathbf{R}_\mathrm{s}$ and update $\mathbf{R}$. Based on Eq. 9, let

$$\mathbf{H} \overset{\triangle}{=} \mathbf{D} - \lambda_s(2\mathbf{U}_\mathrm{s}^\mathrm{T} - \mathbf{1}_{\mathbf{T}_\mathrm{s}})^\mathrm{T}\mathbf{R}_\mathrm{s}\mathbf{Q}_\mathrm{s}^\mathrm{T},$$

We get another representation of $\mathbf{J}$:

$$\mathbf{J} = \underbrace{tr(\mathbf{R}^\mathrm{T}\mathbf{H})}_{\mathbf{J}_3} + \underbrace{\lambda_s tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}})}_{\mathbf{J}_4}, \tag{12}$$

where $\mathbf{J}_4 = \lambda_s tr(\mathbf{R}_\mathrm{s}^\mathrm{T}\mathbf{U}_\mathrm{s}^\mathrm{T}\mathbf{1}_{\mathbf{U}_\mathrm{s}})$ is a constant given $\mathbf{R}_\mathrm{s}$ and $\mathbf{U}_\mathrm{s}$. Therefore, only $\mathbf{J}_3$ needs to be minimized. We update $\mathbf{R}$ to minimize $\mathbf{J}_3$ as follows, $\forall\ n = 1, \ldots N$:

$$r_{mn} = \begin{cases} 1 & m = \arg\min_k h_{kn} \\ 0 & otherwise, \end{cases} \tag{13}$$

where h_{kn} is the element of $\mathbf{H}$ and r_{mn} is the element of $\mathbf{R}$.

The above E-step itself is an EM-like process because we need to update $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$ iteratively until $\mathbf{J}$ converges. The objective function $\mathbf{J}$ decreases monotonically at each step.

M-step:

After knowing the labels of visual primitives and spatial context groups ($\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$), we want to estimate better visual lexicons $\boldsymbol{\Omega}$ and Ψ_s. From Eq. 3, $\mathbf{D}$ and $\mathbf{D}_\mathrm{s}$ are not interlaced, thus $\mathbf{U}$ and $\mathbf{U}_\mathrm{s}$ can be optimized separately. We apply the following two steps to update $\mathbf{U}$ and $\mathbf{U}_\mathrm{s}$ separately.

1) Recalculate the cluster centroid for each feature word class $\{\mathbf{u}_m\}_{m=1}^M$ like traditional k-means algorithm, with Euclidean distance. Update $\mathbf{U}$ and $\mathbf{D}$ to decrease $\mathbf{J}$.
2) Recalculate the cluster centroid for each spatial pattern class $\{\mathbf{u}_m^{(\mathrm{s})}\}_{m=1}^{M_s}$, with Hamming distance. Update $\mathbf{U}_\mathrm{s}$ and $\mathbf{D}_\mathrm{s}$ to decrease $\mathbf{J}$.

Both of the above steps guarantee that $\mathbf{J}$ is decreasing, therefore the whole M-step decreases $\mathbf{J}$ monotonically. Our method is actually a nested-EM algorithm because there are two nested EM processes, where the E-step itself is an EM process that supports a bottom-up/top-down update between labels of low level feature words and high level spatial patterns. We describe this clustering approach in Algorithm 1.

Because the solution spaces of $\mathbf{R}$ and $\mathbf{R}_\mathrm{s}$ are discrete and finite, according to the monotonic decreasing of $\mathbf{J}$ at each step of spatial context-aware clustering, we have theorem 1.

Theorem 1 (Convergence). *The spatial context-aware clustering algorithm in Algorithm 1 can converge in finite steps.*

D. Simulations

To illustrate our spatial context-aware clustering, we synthesize a spatial dataset. A concrete example of this spatial

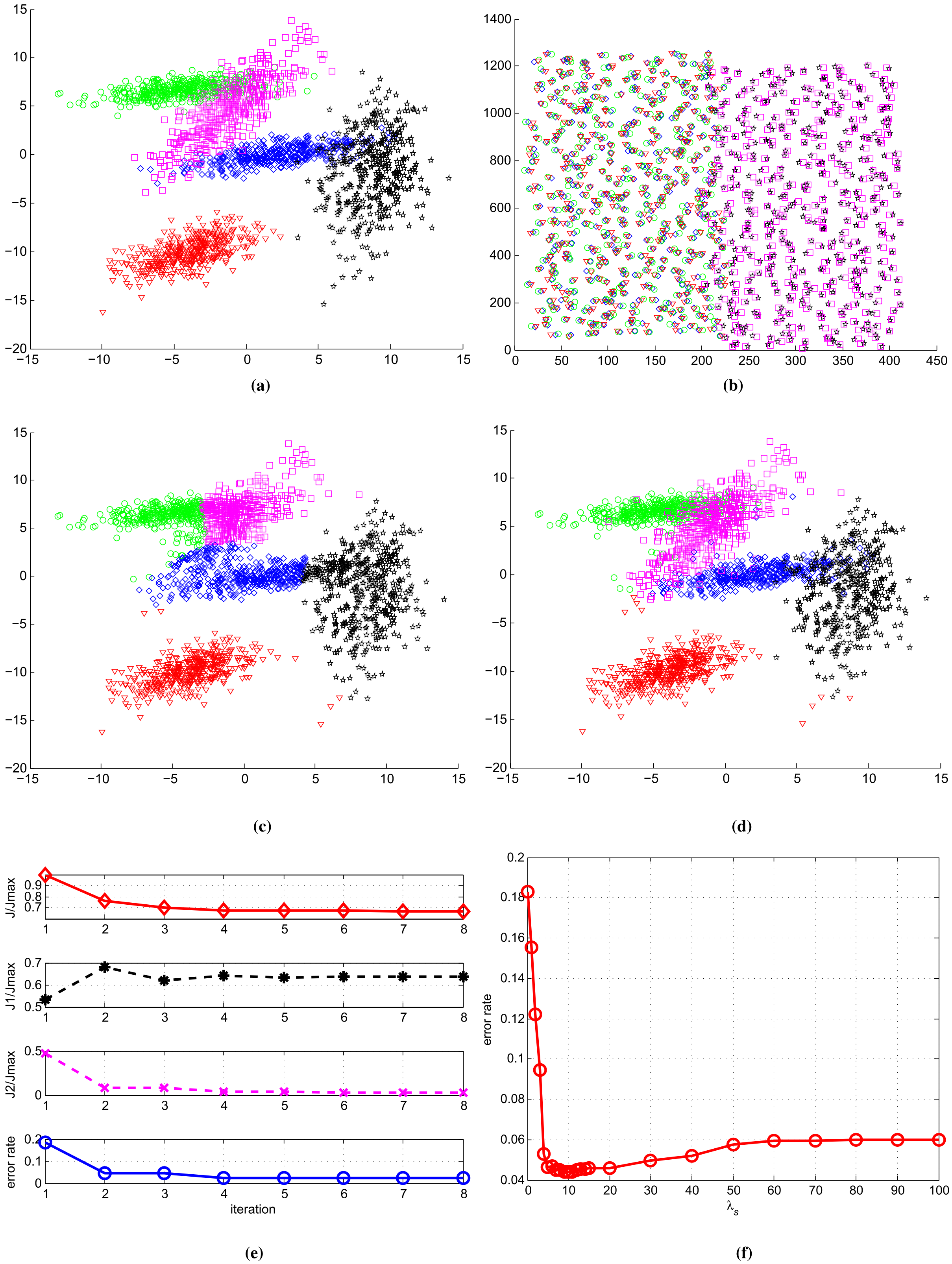


Fig. 2. Context-aware clustering on the synthesized spatial data and the comparison with the k-means algorithm. Parameter used are $M = 5$, $M_S = 2$ and $\epsilon = 100$ in searching for ϵ-NN spatial groups. See texts for descriptions. Best seen in color. (a) Feature Domain. (b) Spatial Domain. (c) k-means Clustering ($k = 5$). (d) Context-Aware Clustering ($\lambda_S = 10$). (e) Performance ($\lambda_S = 10$). (f) Error *vs.* λ_S.

dataset can be an image. All the samples have two representations: (1) feature domain, $\mathbf{f} \in \mathbb{R}^2$ and (2) spatial domain $(x, y) \in \mathbb{N} \times \mathbb{N}$ as shown in Fig. 2(a) and (b), respectively. In our case, we have 5 categories of visual primitives labeled as: ∇, ❍, ♢, □ or ☆. In the spatial domain, {☆, □} is generated together to form a co-occurrent contextual pattern, while {∇, ❍, ♢} is the other visual pattern. In the feature domain, each of the 5 categories has 400 samples and are generated based on Gaussian distributions of different means and variances. Based on the feature domain only, clustering is a challenging problem because some of these Gaussian distributions are heavily overlapped. For example, there is a heavy overlap between categories '❍' and '□,' '□' and ♢, ♢ and ☆. Our tasks are (1) clustering visual primitives into feature words, and (2) recover the spatial patterns $\mathcal{P}_1 = \{$☆, □$\}$ and $\mathcal{P}_2 = \{$∇, ❍, ♢$\}$.

We compare the performances of the spatial context-aware clustering with different choices of λ_{s} ($\lambda_{\mathrm{s}} = 0, 10$) in Fig. 2(c) and (d), where $\lambda_{\mathrm{s}} = 0$ gives the same results as the k-means clustering. The major differences of the clustering results appear from the overlapping categories '❍,' '□,' ♢ and ☆. Although they are heavily overlapped with each other, most of the samples are still correctly labeled with the help of their spatial contexts. For example, although it is difficult to determine a sample υ located in the overlapped regions of '♢' and '☆' in the feature space, we can resolve the ambiguity by observing the spatial contexts of υ. If a '∇' or a ❍ is found in its spatial contexts, then υ should be labeled as '♢' because the discovered visual pattern {∇, ❍, ♢} supports such a label.

Fig. 2(e) shows the iterations of our algorithm with $\lambda_{\mathrm{s}} = 10$. Each iteration corresponds to an individual E-step or an M-step until converge. We decompose the objective function into $\mathbf{J} = \mathbf{J}_1 + \mathbf{J}_2$, where $\mathbf{J}$, $\mathbf{J}_1$, $\mathbf{J}_2$ are the red, black and pink curves respectively. All these three curves are normalized by $\mathbf{J}_{max} = \mathbf{J}^0$, which is the $\mathbf{J}$ value at the initialization step. Compared to the k-means clustering which minimizes distortions $\mathbf{J}_1$ in feature space only, our context-aware clustering sacrifices $\mathbf{J}_1$ to gain larger decrease of distortion $\mathbf{J}_2$ in the spatial context space, which gives a smaller total distortion $\mathbf{J}$. The error rate curve (blue) describes the percentage of samples that are wrongly labeled at each step, and we notice that it decreases consistently with our objective function $\mathbf{J}$.

In terms of clustering errors shown in Fig. 2(f), the context-aware clustering ($error\ rate \leq 6.0\%$ when $\lambda_s \geq 5$) performs significantly better than the k-means method ($error\ rate = 18.3\%$). The parameter λ_s balances the two clustering criteria: (1) clustering based on visual features $\mathbf{f}$ ($\mathbf{J}_1$) and (2) clustering based on spatial contexts ($\mathbf{J}_2$). The smaller the λ_{s}, the more faithful the clustering results follow the feature space, where samples have similar features are grouped together. An extreme case is $\lambda_{\mathrm{s}} = 0$ when no regularization is applied in Eq. 3 by ignoring the feedback from contexts. In such a case, our context-aware clustering is equal to k-means clustering. On the other hand, a larger λ_s favors the clustering results that support the discovered context patterns, thus samples have similar contexts are more likely to be grouped together. For this case, the value of λ_{s} with the best balance is 10. As shown in Fig. 2(f), $\lambda_{\mathrm{s}} = 10$ achieves the lowest clustering error rate: 4.4%.

It is worth noting that the above experiments are all performed under the same k-means initialization. To evaluate the stability of our method, we run k-means 100 times with different initializations. Our context-aware clustering (with $\lambda_s = 10$) has an average error rate 4.24% with small standard deviation 0.17%. It shows that our method is not sensitive to the initialization of k-means.

IV. Multi-Context-Aware Clustering

A. Problem Statement

Besides spatial context, feature context is also important for visual pattern discovery. We will propose in this section the multi-context-aware clustering that utilizes both feature and spatial contexts. In this multi-context-aware clustering, each visual primitive $\upsilon_n \in \mathcal{D}_\upsilon$ is characterized by c types of features: $\upsilon_n = \{\mathbf{f}_n^{(i)}\}_{i=1}^c$, where $\mathbf{f}_n^{(i)} \in \mathbb{R}^{d_i}$. For example, an image can be represented by color, shape, texture and any other visual features. These features of υ_n correspond to a feature context group $\mathcal{G}_n^{(\mathrm{v})}$.

By k-means clustering, each type of features $\{\mathbf{f}_n^{(i)}\}_{n=1}^N$ can produce a feature word lexicon $\mathbf{\Omega}_i$ ($|\mathbf{\Omega}_i| = M_i$). Each $\upsilon_n \in \mathcal{D}_\upsilon$ then generates a feature context transaction $\mathbf{t}_n^{(\mathrm{v})} \in \mathbb{R}^{\sum_{i=1}^c M_i}$ to represent $\mathcal{G}_n^{(\mathrm{v})}$, which is defined in the following.

Definition 2 (Feature context transaction). *The feature context transaction of data υ_n refers to the co-occurrences of multi-view feature words in the feature context group of υ_n.*

Using label indicator matrices $\{\mathbf{R}_i\}_{i=1}^c$ obtained from the c types of features, we can represent the feature context transaction database as a binary matrix:

$$\mathbf{T}_{\mathrm{v}} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \vdots \\ \mathbf{R}_c \end{bmatrix}, \tag{14}$$

where $\mathbf{R}_i \in \mathbb{R}^{M_i \times N}$, and the binary entry $r_{mn}^{(i)} = 1$ only if υ_n is labeled with the m^{th} discovered feature word based on the i^{th} types of features $\{\mathbf{f}_n^{(i)}\}_{n=1}^N$; $\mathbf{T}_{\mathrm{v}} \in \mathbb{R}^{\sum_{i=1}^c M_i \times N}$, and the n^{th} column of $\mathbf{T}_{\mathrm{v}}$ is just the feature context transaction of data υ_n, *i.e.*, $\mathbf{t}_n^{(\mathrm{v})}$.

After clustering these N feature context transactions, the data points can be labeled by a high level feature pattern lexicon Ψ_{v} ($|\Psi_{\mathrm{v}}| = M_{\mathrm{v}}$), which partition the given data in $\mathcal{D}_{\mathrm{v}}$ using multiple features. Besides feature contexts, following the analysis of Sec. III, we can further explore the spatial dependencies among primitives to find a higher level spatial pattern lexicon Ψ_{s} ($|\Psi_{\mathrm{s}}| = M_{\mathrm{s}}$). On the other hand, once we find spatial patterns, we can use them to tune the primitive clustering. Through such a top down refinement, spatial patterns can help to improve feature pattern constructions. Afterwards, 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. The idea described above is shown in Fig. 3 using three types of features. To achieve this objective, we propose a regularized

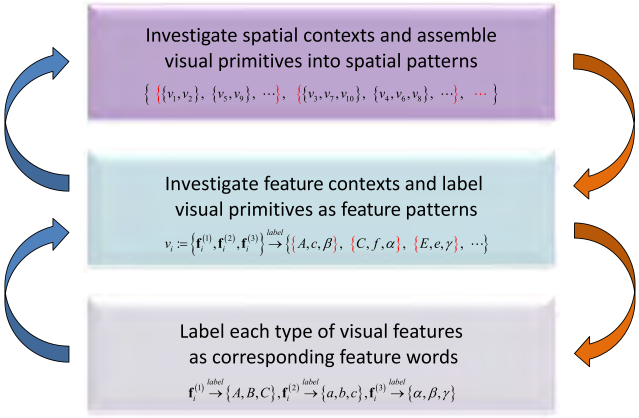

Investigate spatial contexts and assemble visual primitives into spatial patterns
$\{ \{\{v_1, v_2\}, \{v_5, v_9\}, \cdots\}, \{\{v_3, v_7, v_{10}\}, \{v_4, v_6, v_8\}, \cdots\}, \cdots \}$
Investigate feature contexts and label visual primitives as feature patterns
$v_i := \{\mathbf{f}_i^{(1)}, \mathbf{f}_i^{(2)}, \mathbf{f}_i^{(3)}\} \xrightarrow{label} \{\{A, c, \beta\}, \{C, f, \alpha\}, \{E, e, \gamma\}, \cdots\}$
Label each type of visual features as corresponding feature words
$\mathbf{f}_i^{(1)} \xrightarrow{label} \{A, B, C\}, \mathbf{f}_i^{(2)} \xrightarrow{label} \{a, b, c\}, \mathbf{f}_i^{(3)} \xrightarrow{label} \{\alpha, \beta, \gamma\}$

Fig. 3. Multi-context-aware clustering via a self-learning procedure between low-level feature clustering and high-level pattern discovery.

k-means with feature contexts and spatial contexts together:

$$\begin{aligned} \mathbf{J} &= \sum_{i=1}^{c} \sum_{m=1}^{M_i} \sum_{n=1}^{N} r_{mn}^{(i)} \left\| \mathbf{u}_m^{(i)} - \mathbf{f}_n^{(i)} \right\|^2 \\ &\quad + \lambda_{\text{v}} \sum_{m=1}^{M_{\text{v}}} \sum_{n=1}^{N} r_{mn}^{(\text{v})} d_H \left(\mathbf{u}_m^{(\text{v})}, \mathbf{t}_n^{(\text{v})} \right) \\ &\quad + \lambda_{\text{s}} \sum_{m=1}^{M_{\text{s}}} \sum_{n=1}^{N} r_{mn}^{(\text{s})} d_H \left(\mathbf{u}_m^{(\text{s})}, \mathbf{t}_n^{(\text{s})} \right) \\ &= \sum_{i=1}^{c} tr \left(\mathbf{R}_i^{\text{T}} \mathbf{D}_i \right) + \lambda_{\text{v}} tr \left(\mathbf{R}_{\text{v}}^{\text{T}} \mathbf{D}_{\text{v}} \right) + \lambda_{s} tr \left(\mathbf{R}_{\text{s}}^{\text{T}} \mathbf{D}_{\text{s}} \right) \\ &= \underbrace{tr \left(\mathfrak{R}^{\text{T}} \mathfrak{D} \right)}_{\mathbf{J}_\alpha} + \underbrace{\lambda_{\text{v}} tr \left(\mathbf{R}_{\text{v}}^{\text{T}} \mathbf{D}_{\text{v}} \right)}_{\mathbf{J}_\beta} + \underbrace{\lambda_{\text{s}} tr \left(\mathbf{R}_{\text{s}}^{\text{T}} \mathbf{D}_{\text{s}} \right)}_{\mathbf{J}_\gamma}, \end{aligned} \tag{15}$$

where,

- $\lambda_{\text{v}} > 0$ and $\lambda_{\text{s}} > 0$ are constants for regularization;
- $\mathbf{J}_\alpha$ is the total quantization distortions of multiple types of features, in which
  - $\{\mathbf{u}_m^{(i)}\}_{m=1}^{M_i}$ denote M_i quantized feature words based on the i_{th} type of features $\{\mathbf{f}_n^{(i)}\}_{n=1}^{N}$, and form a feature word matrix $\mathbf{U}_i \in \mathbb{R}^{d_i \times M_i}$;
  - $\mathbf{R}_i \in \mathbb{R}^{M_i \times N}$ is a binary label indicator matrix, the entry $r_{mn}^{(i)} = 1$ only if $\mathbf{f}_n^{(i)}$ is labeled with the m^{th} discovered feature word $\mathbf{u}_m^{(i)}$ based on the i^{th} type of features $\{\mathbf{f}_n^{(i)}\}_{n=1}^{N}$;
  - $\mathbf{D}_i \in \mathbb{R}^{M_i \times N}$ denotes a distortion matrix, the entry of its m^{th} row and n^{th} column is the Euclidean square distortion between $\mathbf{u}_m^{(i)}$ and $\mathbf{f}_n^{(i)}$;
  - $\mathfrak{R}$ and $\mathfrak{D}$ are block diagonal matrices from $\{\mathbf{R}_i\}_{i=1}^{c}$ and $\{\mathbf{D}_i\}_{i=1}^{c}$;
- $\mathbf{J}_\beta$ is the quantization distortion of feature context transactions, in which
  - $\{\mathbf{u}_m^{(\text{v})}\}_{m=1}^{M_{\text{v}}}$ denote M_{v} quantized feature patterns, forming a feature pattern matrix $\mathbf{U}_{\text{v}} \in \mathbb{R}^{\sum_{j=1}^{c} M_j \times M_{\text{v}}}$;
  - $\mathbf{R}_{\text{v}} \in \mathbb{R}^{M_{\text{v}} \times N}$ is a binary label indicator matrix, the entry $r_{mn}^{(\text{v})} = 1$ only if υ_n is included the m^{th} discovered feature pattern $\mathbf{u}_m^{(\text{v})}$;
  - $\mathbf{D}_{\text{v}} \in \mathbb{R}^{M_{\text{v}} \times N}$ denotes a distortion matrix, the entry of its m^{th} row and n^{th} column is the Hamming distortion between $\mathbf{u}_m^{(\text{v})}$ and $\mathbf{t}_n^{(\text{v})}$;
- $\mathbf{J}_\gamma$ is the quantization distortion of spatial context transactions, similar to the constrained term $\mathbf{J}_2$ in Eq. 3. So we abuse the symbols between $\mathbf{J}_\gamma$ and $\mathbf{J}_2$. The only difference between them resides in the dimensionality of a spatial context transaction $\mathbf{t}_n^{(\text{s})}$ ($\in \mathbb{R}^{M_{\text{v}}}$) or a prototype of spatial pattern $\mathbf{u}_m^{(\text{s})}$ ($\in \mathbb{R}^{M_{\text{v}}}$), because the number of visual primitive categories here becomes M_{v}, *i.e.*, the number of feature patterns. So we also have the spatial context group matrix $\mathbf{T}_{\text{s}} \in \mathbb{R}^{M_{\text{v}} \times N}$; and spatial pattern matrix $\mathbf{U}_{\text{s}} \in \mathbb{R}^{M_{\text{v}} \times M_{\text{s}}}$.

Because each data point has c types of features, the feature context relations can be represented as a concatenated matrix $\mathbf{Q}_{\text{v}} \in \mathbb{R}^{cN \times N}$ from c identity matrices of size $N \times N$, the following equation is hold:

$$\mathbf{T}_{\text{v}} = \mathfrak{R} \mathbf{Q}_{\text{v}}. \tag{16}$$

Besides, similar to Eq. 7, according to the local spatial neighbor relation matrix of the primitive collection $\mathcal{D}_\upsilon$, *i.e.*, an $N \times N$ matrix $\mathbf{Q}_{\text{s}}$ whose entry $q_{ij} = 1$ only if υ_i and υ_j are local spatial neighbors, we can represent $\mathbf{T}_{\text{s}}$ as

$$\mathbf{T}_{\text{s}} = \mathbf{R}_{\text{v}} \mathbf{Q}_{\text{s}}. \tag{17}$$

Comparing with Eq. 6, $\mathbf{D}_\upsilon$ can be represented by

$$\begin{aligned} \mathbf{D}_{\text{v}} &= -2\mathbf{U}_{\text{v}}^{\text{T}} \mathbf{T}_{\text{v}} + \mathbf{1}_{\mathbf{T}_{\text{v}}} \mathbf{T}_{\text{v}} + \mathbf{U}_{\text{v}}^{\text{T}} \mathbf{1}_{\mathbf{U}_{\text{v}}} \\ &= -2\mathbf{U}_{\text{v}}^{\text{T}} \mathfrak{R} \mathbf{Q}_{\text{v}} + \mathbf{1}_{\mathbf{T}_{\text{v}}} \mathfrak{R} \mathbf{Q}_{\text{v}} + \mathbf{U}_{\text{v}}^{\text{T}} \mathbf{1}_{\mathbf{U}_{\text{v}}}, \end{aligned} \tag{18}$$

where $\mathbf{1}_{\mathbf{T}_{\text{v}}}$ is an $M \times \sum_{i=1}^{c} M_i$ all 1 matrix, and $\mathbf{1}_{\mathbf{U}_{\text{v}}}$ is a $\sum_{i=1}^{c} M_i \times N$ all 1 matrix .

In a similar way, the Hamming distortions $\mathbf{D}_{\text{s}}$ can be formulated as

$$\begin{aligned} \mathbf{D}_{\text{s}} &= -2\mathbf{U}_{\text{s}}^{\text{T}} \mathbf{T}_{\text{s}} + \mathbf{1}_{\mathbf{T}_{\text{s}}} \mathbf{T}_{\text{s}} + \mathbf{U}_{\text{s}}^{\text{T}} \mathbf{1}_{\mathbf{U}_{\text{s}}} \\ &= -2\mathbf{U}_{\text{s}}^{\text{T}} \mathbf{R}_{\text{v}} \mathbf{Q}_{\text{s}} + \mathbf{1}_{\mathbf{T}_{\text{s}}} \mathbf{R}_{\text{v}} \mathbf{Q}_{\text{s}} + \mathbf{U}_{\text{s}}^{\text{T}} \mathbf{1}_{\mathbf{U}_{\text{s}}}, \end{aligned} \tag{19}$$

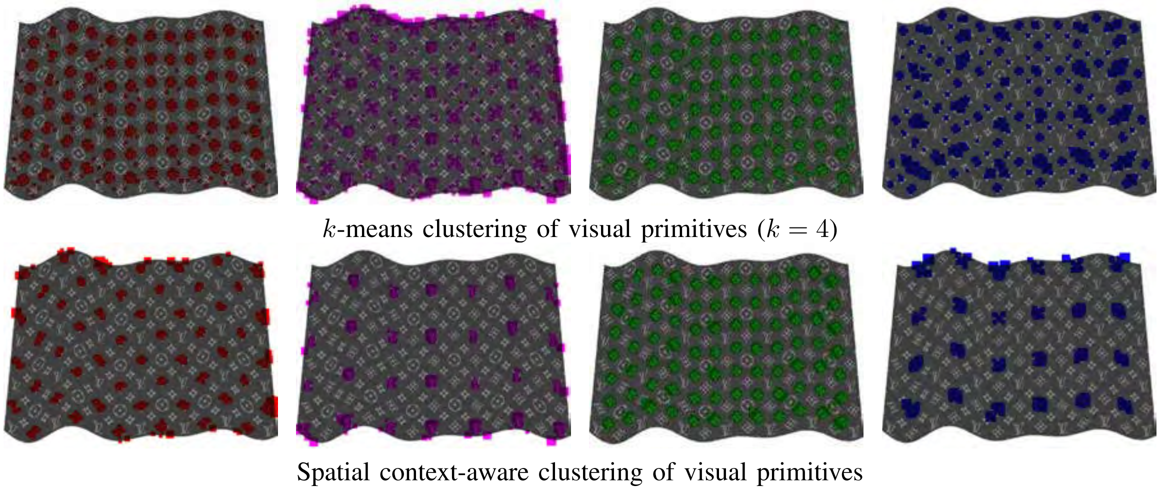
where $\mathbf{1}_{\mathbf{T}_{\text{s}}}$ is an $M_s \times M$ all 1 matrix, and $\mathbf{1}_{\mathbf{U}_{\text{s}}}$ is an $M \times N$ all 1 matrix.

Similar to the spatial context-aware clustering in Sec. III, $\mathbf{J}_\alpha$, $\mathbf{J}_\beta$ and $\mathbf{J}_\gamma$ are correlated among each other. We thus cannot minimize $\mathbf{J}$ by minimizing the three terms separately, which makes the objective function of Eq. 15 a challenge. We will in Sec. IV-B show how to decouple the dependencies among them and propose our algorithm to solve this optimization function.

B. Algorithm

Since there are correlations among unknown variables simultaneously in Eq. 15, we cannot estimate them simultaneously. So we have to decouple the dependencies among the terms of Eq. 15. Above all, we initialize feature words, feature patterns and spatial patterns gradually by k-means. Next, we can take each of $\mathbf{R}_{\text{v}}$, $\mathbf{R}$, and $\mathbf{R}_{\text{s}}$ as the common factor for extraction. We derive Eq. 15 as:

$$\begin{aligned} &\mathbf{J}(\mathfrak{R}, \mathbf{R}_{\text{v}}, \mathbf{R}_{\text{s}}, \mathfrak{D}, \mathbf{D}_{\text{v}}, \mathbf{D}_{\text{s}}) \\ &\quad = tr(\mathbf{R}_{\text{v}}^{\text{T}} \mathbf{H}_{\text{v}}) + tr(\mathfrak{R}^{\text{T}} \mathfrak{D}) + \lambda_{\text{s}} tr(\mathbf{R}_{\text{s}}^{\text{T}} \mathbf{U}_{s}^{\text{T}} \mathbf{1}_{U_{\text{s}}}) \qquad (20) \\ &\quad = tr(\mathfrak{R}^{\text{T}} \mathfrak{H}) + \lambda_{\text{s}} tr(\mathbf{R}_{\text{s}}^{\text{T}} \mathbf{D}_{\text{s}}) + \lambda_{\text{v}} tr(\mathbf{R}_{\text{v}}^{\text{T}} \mathbf{U}_{\text{v}}^{\text{T}} \mathbf{1}_{U_{\text{v}}}) \qquad (21) \\ &\quad = tr(\mathbf{R}_{\text{s}}^{\text{T}} \mathbf{H}_{\text{s}}) + tr(\mathfrak{R}^{\text{T}} \mathfrak{D}) + \lambda_{\text{v}} tr(\mathbf{R}_{\text{v}}^{\text{T}} \mathbf{D}_{\text{v}}), \qquad (22) \end{aligned}$$



k-means clustering of visual primitives ($k = 4$)

Spatial context-aware clustering of visual primitives

Fig. 4. Pattern discovery from a mono-colored LV monogram picture. Best seen in color. (a) k-means clustering of visual primitives ($k = 4$). (b) Spatial context-aware clustering of visual primitives.

where

$$\mathbf{H}_\text{v} = \lambda_\text{v}\mathbf{D}_\text{v} - \lambda_\text{s}(2\mathbf{U}_\text{s}^\text{T} - \mathbf{1}_{\mathbf{T}_\text{s}})^\text{T}\mathbf{R}_\text{s}\mathbf{Q}_\text{s}^\text{T}, \tag{23}$$

$$\mathfrak{H} = \mathfrak{D} - \lambda_\text{v}(2\mathbf{U}_\text{v}^\text{T} - \mathbf{1}_{\mathbf{T}_\text{v}})^\text{T}\mathbf{R}_\text{v}\mathbf{Q}_\text{v}^\text{T}, \tag{24}$$

$$\mathbf{H}_\text{s} = \lambda_s\mathbf{D}_\text{s}, \tag{25}$$

where the size of $\mathbf{H}_\text{v}$, $\mathfrak{H}$ and $\mathbf{H}_\text{s}$ are $M \times N$, $\sum_{i=1}^{c} M_i \times cN$ and $M_\text{s} \times N$, and $\mathbf{H}$ contains c diagonal blocks $\{\mathbf{H}_i\}_{i=1}^{c}$.

We can successively update the three label indicator matrices $\mathbf{R}_\text{v}$, $\mathfrak{R}$, and $\mathbf{R}_\text{s}$ when the cluster centroid matrices $\mathbf{U}_\text{v}$, $\{\mathbf{U}_i\}_{i=1}^{c}$, and $\mathbf{U}_\text{s}$ are fixed. To minimize Eq. 15, the following label indicator matrices update criteria will be adopted, $\forall n = 1, 2, \ldots, N$,

$$r_{mn}^{(\text{v})} = \begin{cases} 1 & m = \arg\min_k h_{kn}^{(\text{v})} \\ 0 & otherwise, \end{cases} \tag{26}$$

$$r_{mn}^{(j)} = \begin{cases} 1 & m = \arg\min_k h_{kn}^{(j)} \\ 0 & otherwise, \end{cases} \tag{27}$$

$$r_{mn}^{(\text{s})} = \begin{cases} 1 & m = \arg\min_k h_{kn}^{(\text{s})} \\ 0 & otherwise, \end{cases} \tag{28}$$

where $h_{kn}^{(\text{v})}$, $r_{mn}^{(\text{v})}$, $h_{kn}^{(i)}$, $r_{mn}^{(i)}$, $h_{kn}^{(\text{s})}$ and $r_{mn}^{(\text{s})}$ are the entries of $\mathbf{H}_\text{v}$, $\mathbf{R}_\text{v}$, $\mathbf{H}_i$, $\mathbf{R}_i$, $\mathbf{H}_\text{s}$ and $\mathbf{R}_\text{s}$, respectively. As long as the objective function $\mathbf{J}$ defined by Eq. 15 is decreasing, $\mathbf{R}_\upsilon$ and $\mathbf{R}$ can be continually refined, followed by the bottom-up updates of $\mathbf{R}_\text{v}$ and $\mathbf{R}_\text{s}$.

Furthermore, provided the label indicator matrices $\mathbf{R}_\text{v}$, $\mathbf{R}$, and $\mathbf{R}_\text{s}$, the corresponding centroid matrices $\mathbf{U}_\text{v}$, $\{\mathbf{U}_i\}_{i=1}^{c}$, and $\mathbf{U}_\text{s}$ can be updated, and so as the corresponding distortion matrices $\mathbf{D}_\text{v}$, $\{\mathbf{D}_i\}_{i=1}^{c}$, and $\mathbf{D}_\text{s}$, which will also make the objective function of Eq. 15 decrease.

We propose an iterative algorithm of multi-context-aware clustering in Algorithm 2. Similar to Algorithm 1, this Algorithm is also convergent since the solution spaces of $\mathfrak{R}$, $\mathbf{R}_\text{v}$, and $\mathbf{R}_\text{s}$ are discrete and finite, and the objective function defined by Eq. 15 is monotonic decreasing at each step. Clearly, our multi-context-aware clustering will be degenerated to the spatial context-aware clustering if there is only one type of feature and we set $\lambda_\text{v} = 0$ in Eq. 15 to remove the $\mathbf{J}_\beta$ term.

Algorithm 2 Multi-Context-Aware Clustering

```
input  : database $\mathcal{D}_v = \{v_n\}_{n=1}^{N}$;
         feature neighbor relations $\mathbf{Q}_\text{v}$;
         spatial neighbor relations $\mathbf{Q}_\text{s}$;
         parameters: $\{M_i\}_{i=1}^{c}$, $M_\text{v}$, $M_\text{s}$, $\lambda_\text{v}$, $\lambda_\text{s}$
output : feature word lexicons: $\{\mathbf{\Omega}_i\}_{i=1}^{c}$ ($\{\mathbf{U}_i\}_{i=1}^{c}$); feature
         pattern lexicon: $\mathbf{\Psi}_\text{v}$ ($\mathbf{U}_\text{v}$); spatial pattern lexicon: $\mathbf{\Psi}_\text{s}$
         ($\mathbf{U}_\text{s}$); clustering results $\{\mathbf{R}_i\}_{i=1}^{c}$, $\mathbf{R}_\text{v}$, $\mathbf{R}_\text{s}$
Init: perform $k$-means clustering from bottom up to obtain
      $\{\mathbf{U}_i\}_{i=1}^{c}$, $\mathbf{U}_\text{v}$, $\mathbf{U}_\text{s}$;
while not converged do
    R-step: fix $\{\mathbf{U}_i\}_{i=1}^{c}$, $\mathbf{U}_\text{v}$, $\mathbf{U}_\text{s}$, recursively top-down /
            bottom-up update $\{\mathbf{R}_i\}_{i=1}^{c}$, $\mathbf{R}_\text{v}$, $\mathbf{R}_\text{s}$
    if J is decreasing then
        goto R-setp
    else
        goto D-step
    D-step: fix $\{\mathbf{R}_i\}_{i=1}^{c}$, $\mathbf{R}_\text{v}$, $\mathbf{R}_\text{s}$, update $\{\mathbf{U}_i\}_{i=1}^{c}$, $\mathbf{U}_\text{v}$, $\mathbf{U}_\text{s}$
return $\{\mathbf{U}_i\}_{i=1}^{c}$, $\mathbf{U}_\text{v}$, $\mathbf{U}_\text{s}$, $\{\mathbf{R}_i\}_{i=1}^{c}$, $\mathbf{R}_\text{v}$, $\mathbf{R}_\text{s}$.
```

V. Experiments

In the following experiments, we set $M_i = M_\text{v}$, $\forall i = 1, 2, \ldots, c$ in multi-context-aware clustering. Besides, to help parameter tuning, we let $\lambda_\text{s} = \tau_\text{s}|\mathbf{J}_1^0/\mathbf{J}_2^0|$ in spatial context-aware clustering, $\lambda_\text{v} = \tau_\text{v}|\mathbf{J}_\alpha^0/\mathbf{J}_\beta^0|$ and $\lambda_\text{s} = \tau_\text{s}|\mathbf{J}_\alpha^0/\mathbf{J}_\gamma^0|$ in multi-context-aware clustering, where $\mathbf{J}_x^0$ ($x = 1, 2, \alpha, \beta, \gamma$) is the initial value of $\mathbf{J}_x$ defined by Eq. 3, or Eq. 15, and the nonnegative constants τ_v and τ_s are the auxiliary parameters to balance the influences from feature co-occurrences and spatial co-occurrences, respectively.

A. Spatial Visual Pattern Discovery

To validate whether our methods can really capture spatial visual patterns [21], we test a number of images using spatial context-aware clustering and/or multi-context-aware clustering presented in Figs. 4 and 5. Given an image, we first extract one or more (*e.g.*, c types of) features for the detected visual primitives: $\mathcal{D}_\upsilon = \{\upsilon_n\}_{n=1}^{N}$, and apply spatial K-NN groups to build spatial context group database $\{\mathcal{G}_n^{(\text{s})}\}_{n=1}^{N}$.

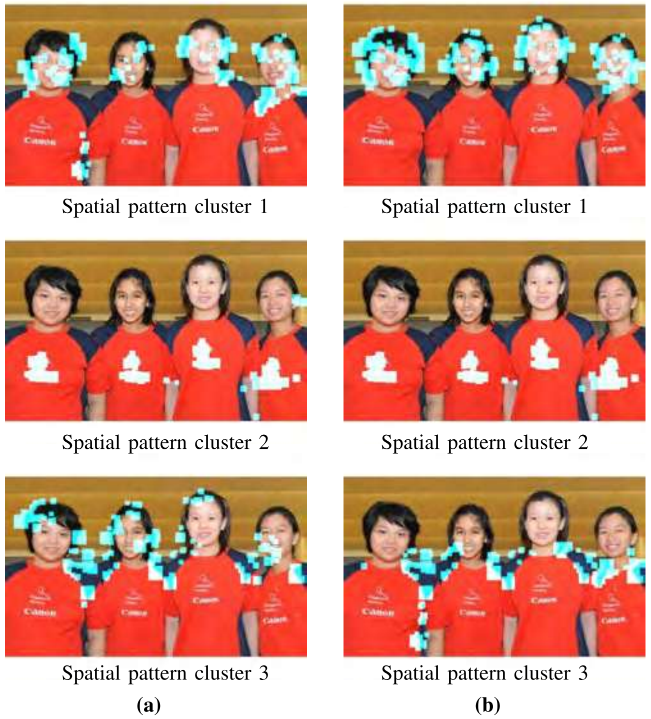

Spatial pattern cluster 1
Spatial pattern cluster 1
Spatial pattern cluster 2
Spatial pattern cluster 2
Spatial pattern cluster 3
Spatial pattern cluster 3
(a)
(b)

Fig. 5. Pattern discovery from a colored group photo. Best seen in color. (a) Spatial context-aware clustering. (b) Multi-context-aware clustering.

A mono-colored LV monogram fabric image is shown in Fig. 4. 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 [48]. To build spatial context groups, K-NN with $K = 8$ is applied. Other parameters are set as $M = 20$, $M_{\mathrm{s}} = 4$, $\tau_{\mathrm{s}} = 1$ for our spatial context-aware clustering. It is interesting to notice that we can also locate the monogram patterns of different spatial structures in Fig. 4, where different colors indicate different (4 in total) discovered spatial patterns. In comparison, without considering spatial dependencies of visual primitives, k-means clustering cannot obtain satisfactory results.

A comparison between spatial context-aware clustering and multi-context-aware clustering is shown in Fig. 5, where 422 image patches [48] are extracted[2]. In spatial context-aware clustering, SIFT features [48] are used to describe these patches. While in multi-context-aware clustering, the patches are represented by SIFT features [48] and Color Histograms (CH) [22]. 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 spatial context-aware clustering; parameters are: $M = 10$, $M_{\mathrm{s}} = 3$, $\tau_{\mathrm{s}} = 0.8$. The results of the 2^{nd} column is based on multi-context-aware clustering; parameters are: $M_i = 10$, $\forall\ i = 1, 2$, $M_{\mathrm{v}} = 10$, $M_{\mathrm{s}} = 3$, $\tau_{\mathrm{v}} = 1.5$, $\tau_{\mathrm{s}} = 0.8$. By using multiple features, the discovered patterns are more accurate. Particularly, in spatial context-aware clustering, there are more confusions between face patterns and edge patterns than those in multi-context-aware clustering.

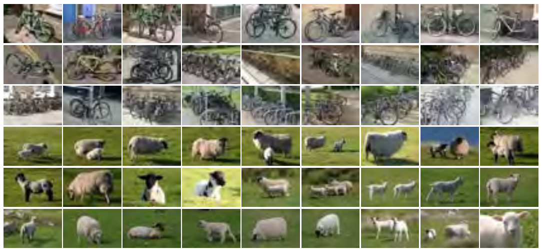


Fig. 6. Two scene categories: "sheep+grass" and "bicycle+road."

TABLE I

CLUSTERING PERFORMANCE OF IMAGE REGIONS FROM IMAGE COLLECTION SHOWN IN FIG. 6

Feature	Method	Error
TH	k-means	34.43%
CH	k-means	41.80%
pHOG	k-means	37.30%
All features	Concatenation for k-means	40.98%
All features	Concatenation with context transactions for k-means	10.65%
All features	**Multi-context-aware**	**5.73%**

B. Image Region Clustering Using Multiple Contexts

To evaluate how much feature contexts and spatial contexts can improve the clustering performance, in this section, we perform image region clustering to test our proposed multi-context-aware clustering on MSRC-V2 dataset [49]. The ground-truth labeling of MSRC-V2 is provided by [50]. To describe each region (*i.e.*, the visual primitive) in MSRC-V2 dataset, we employ three types of features as in [22]: Texton Histogram (TH), Color Histogram (CH), and pyramid of HOG (pHOG) [51]. The dimensionalities of TH, CH, pHOG, are 400, 69, 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 compositions and our goal is to cluster image regions by leveraging the spatial co-occurrence patterns. As baseline methods, k-means clustering results on each single feature type (*TH/CH/pHOG, k-means*) or feature concatenation (*All features, concatenation for k-means*) are compared. Besides that, as context features, feature context transactions and spatial context transactions are concatenated with all feature types for k-means clustering (*All features, concatenation with context transactions for k-means*) to compare with our multi-context-aware clustering on multiple types of features (*All features, multi-context-aware*).

From MSRC-V2 dataset [49], we first select a collection of images with two region pairs that often appear together in an image: "sheep+grass" and "bicycle+road" as shown in Fig. 6. 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."

Table I shows the clustering errors of k-means clustering on individual features and concatenated features, where the best

[2] This image is from source: http://www.abf-online.org/.

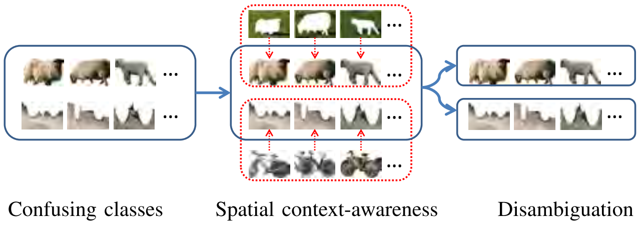


Fig. 7. Class disambiguation by using spatial contexts.

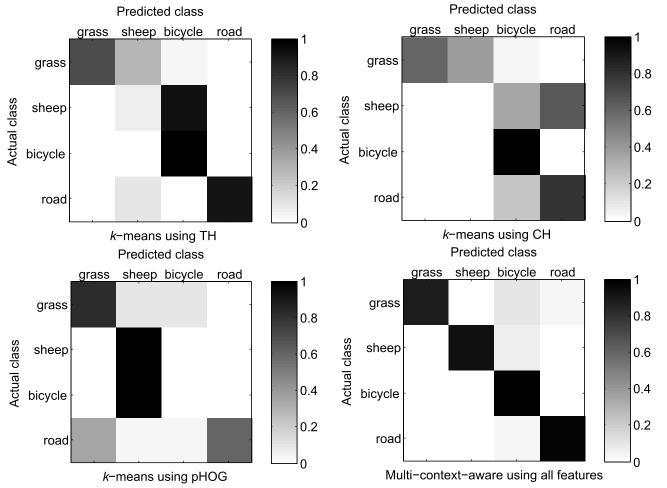


Fig. 8. Confusion matrices of k-means clustering and multi-context-aware clustering for image regions from image collection shown in Fig. 6.

Fig. 9. Sample images of five region compositions: “sheep+grass,” “cow+grass,” “aeroplane+grass+sky,” “boat+water,” and “bicycle+road.”

TABLE II

CLUSTERING PERFORMANCE OF IMAGE REGIONS FROM IMAGE COLLECTION SHOWN IN FIG. 9

Feature	Method	Error
TH	k-means	44.31%
CH	k-means	55.21%
pHOG	k-means	47.63%
All features	Concatenation for k-means	38.39%
All features	Concatenation with context transactions for k-means	33.65%
All features	**Multi-context-aware**	**29.86%**

error rate is 10.65%. In contrast, with suitable parameters, our multi-context-aware clustering can achieve a much better result with error rate 5.73%, which significantly enhances the clustering performance. The parameters used are: $k = 4$ for k-means clustering; and $M_i = 4$, $\forall\ i = 1, 2, 3$, $M_\text{v} = 4$, $M_\text{s} = 2$, $\tau_\text{v} = 3.5$, $\tau_\text{s} = 1$ for multi-context-aware clustering. Each result in Table I is obtained by selecting the best (*i.e.*, the minimum total distortion) from 100 random repetitions.

In the case of using multiple features, k-means clustering on the concatenated features still suffers from the confusion between “sheep” class and “road” class as shown in Fig. 7, where the “sheep” regions are mislabeled as the “road” class. However, by exploring spatial contexts of image regions, our multi-context-aware clustering can 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.

In order to further evaluate how multiple contexts improve the clustering of individual region classes, we show the confusion matrices of k-means clustering and our multi-context-aware clustering in Fig. 8. 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 cause 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, our multi-context-aware clustering can well improve the 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 our multi-context-aware clustering in dealing with image regions of clear spatial contexts. We next evaluate our method on image regions of ambiguous spatial contexts. In this experiment, we focus on images describing frequent occurring region compositions and select 5 different compositions with in total 9 categories of image regions from MSRC-V2 dataset [49]. Some sample images are shown in Fig. 9. In this image collection, there are 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,” “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 multi-context-aware clustering are shown in Table II, where the same 10% seeds per category from ground-truth are randomly chosen for initialization. The clustering error rate of our multi-context-aware clustering 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. In k-means clustering, we set $k = 9$ as there are 9 different

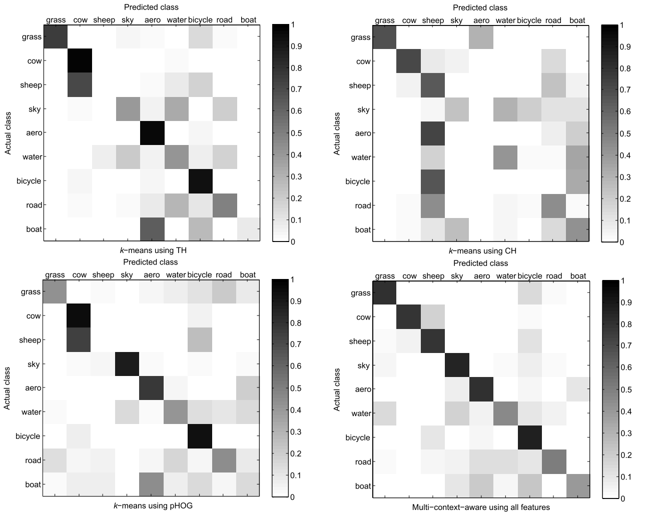


Fig. 10. Confusion matrices of k-means clustering and multi-context-aware clustering for image regions from image collection shown in Fig. 9.

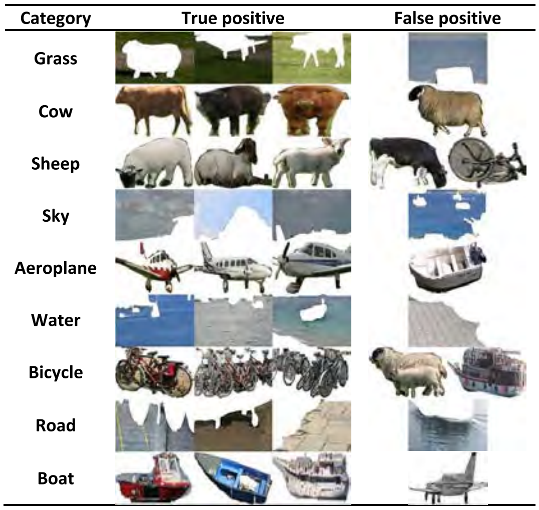


Fig. 11. Exemplar clustering results of multi-context-aware clustering.

types of image regions. Similar to the setting on the dataset shown in Fig. 6, we also set $\tau_{\mathrm{v}} = 3.5$, $\tau_{\mathrm{s}} = 1$ in multi-context-aware clustering. Other parameters used in our approach are $M_i = 9,\ \forall\ i = 1, 2, 3,\ M_{\mathrm{v}} = 9,\ M_{\mathrm{s}} = 5$.

Besides the total clustering error rates, we also compare the clustering performance of individual region classes using k-means clustering and our multi-context-aware clustering in Fig. 10. Because both "sheep" and "cow" share the same spatial context, *i.e.*, "grass," it is difficult to utilize "grass" to distinguish "sheep" and "cow" using the TH and pHOG features. However, our approach can leverage the CH features to distinguish "sheep" and "cow." On the other hand, with the help of spatial context regions, the mislabeled "bicycle" and "aeroplane" regions in CH features and the mislabeled "boat" regions in TH and pHOG features can be partially corrected using our method. Overall, due to the regularization of feature contexts and spatial contexts, our approach can effectively refine the mislabeling results of k-means clustering.

Some representative clustering results of our approach are shown in Fig. 11. Despite large intra-class variations, our method can still obtain a satisfactory clustering results by using both spatial and feature contexts. For example, the "cow" regions are with different colors and perspectives. We also note that in Fig. 9, there contain "water" regions in some "sheep+grass" and "cow+grass" region compositions. These small amount 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.

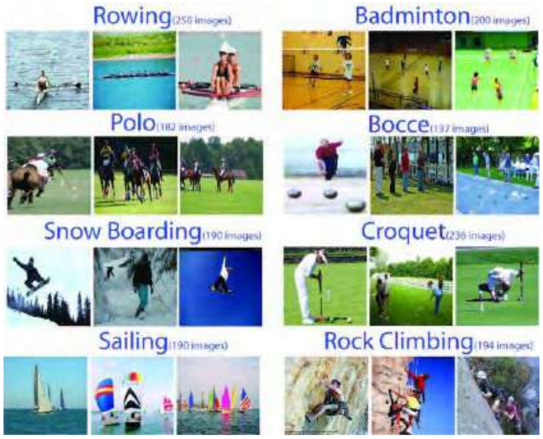


Fig. 12. Sample images of different sports events in UIUC sports dataset.

TABLE III

Classification Accuracy on UIUC Sports Dataset

Feature	BOW (k-means)	BOW (our method)
Accuracy	71.78 ± 1.52%	73.42 ± 1.42%

C. Bag-of-Words for Image Classification

As validated by the simulation experiment in Sec. III-D, our method can build a better visual vocabulary using spatial contexts of visual primitives than k-means clustering. To justify the effectiveness of our method on the real world dataset, we conduct an experiment on the UIUC sports dataset [52]. It contains 8 categories of sport events: rowing, badminton, polo, bocce, snowboarding, croquet, sailing, and rock climbing, with 137 to 250 images in each. Some sample images are shown in Fig. 12. We use MSER detector [53] to generate visual primitives and use SIFT descriptor [48] for primitive representation. Finally, we obtain $4,997,849$ visual primitives totally. After that, we can build a bag-of-word representation (BOW) for each image using k-means clustering or our method. We then train a SVM classifier with RBF kernel using 90 images per category and test on the rest over 5 random training/test splits. In k-means clustering, we set k to 256. In our spatial context-aware clustering, K-NN with $K = 8$ is applied to build spatial context groups. In addition, we set $\tau_{\mathrm{s}} = 0.5$, $M = 256$, and $M_{\mathrm{s}} = 64$. We report the results in Table III. As our method learns visual vocabulary with the spatial configuration

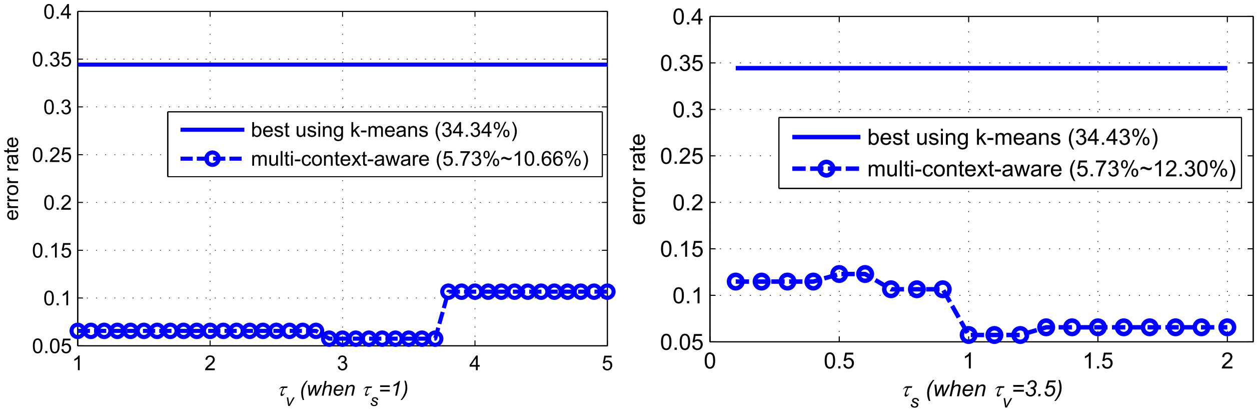


Fig. 13. Clustering error rate of multi-context-aware clustering for image regions from image collection shown in Fig. 6 w.s.t. parameters: τ_{s} and τ_{v}.

information of visual primitives which ignored by k-means clustering, we can see that the BOW features built by our method outperforms than those built by k-means clustering.

D. Parameter Comparison and Selection

We use the image dataset shown in Fig. 6 to study how the two parameters τ_{v} and τ_{s} influence the image region clustering result. It is worth noting that in our multi-context clustering, the larger the τ_{v} and the smaller the τ_{s}, the more faithful the clustering results follow the multiple feature spaces, where image regions of similar features are grouped together; while a smaller τ_{v} and a larger τ_{s} favor the clustering results that support the discovered spatial patterns, thus regions have similar spatial contexts are more likely to be grouped together. To study the impact of parameters, we fix $M_i (i = 1, 2, 3)$, M_{v}, M_{s}, but vary τ_{v} and τ_{s} in turn. We then draw the curves of clustering error rates in Fig. 13. When we fix $\tau_{\mathrm{s}} = 1$ and alter τ_{v} from $[1, 5]$, the clustering error rates fall into $[5.73\%, 10.66\%]$. The best result (*i.e.*, 5.73% error rate) can be obtained when $\tau_{\mathrm{v}} = 3.5$, which balances the distortions between primitive features and visual patterns. Then we fix $\tau_{\mathrm{v}} = 3.5$ and alter τ_{s} from $[0.1, 2.0]$. The largest error rate of 12.3% is still much better than the best one using k-means clustering. The best agreement is reached when $\tau_{\mathrm{s}} = 1$ by balancing the feature domain and spatial domain, where the lowest error rate 5.73% is attained. The experiment also suggests the optimal parameters for this dataset are: $\tau_{\mathrm{v}} = 3.5, \tau_{\mathrm{s}} = 1$, which we have used to report the result in Table I.

Besides the two explicit parameters τ_{v} and τ_{s} studied in Fig. 13, we discuss how to set the other parameters involved in our method. Firstly, similar to the build of visual vocabulary using k-means clustering, it is data dependent to set the feature word number (M), feature pattern number (M_{v}) and spatial pattern number (M_{s}). For example, in image region clustering experiment (Sec. V-B), the numbers of feature words and feature patterns are determined by the region class number, while the number of spatial patterns is determined by the image class number. Secondly, regarding spatial neighborhood size (K) of visual primitives, we notice that a too small or too big K will not provide a satisfactory result. When treating a single image (*e.g.*, images shown in Figs. 4 and 5), the optimal choice of K depends on sizes of the spatial patterns. However, in the image region clustering experiment, we do not need to specify the size of spatial neighborhood, as given a region of an image, all other regions in the same image are considered as in its spatial context group. In such a case, the neighborhood size is automatically set. Moreover, in the image classification experiment (Sec. V-C), it does not require discovering the accurate spatial patterns, but focuses on encoding the spatial configuration of visual primitives into visual vocabulary. Such spatial structure can be learned based on a flexible choice of the sizes of feature words, spatial patterns and spatial neighborhood.

VI. Conclusion

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, image region clustering and image classification validate the advantages of the proposed method.

References

[1] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 284–302, 2010.

[2] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2006, pp. 1605–1614.

[3] Y. Su and F. Jurie, "Visual word disambiguation by semantic contexts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 311–318.

[4] J. Yuan, Y. Wu, and M. Yang, "From frequent itemsets to semantically meaningful visual patterns," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 864–873.

[5] J. Yuan and Y. Wu, "Mining visual collocation patterns via self-supervised subspace learning," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 1–13, Apr. 2012.

[6] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[7] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2664–2677, Sep. 2011.

[8] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 809–816.

[9] J. Yuan and Y. Wu, "Context-aware clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[10] H. Wang, J. Yuan, and Y. Tan, "Combining feature context and spatial context for image pattern discovery," in *Proc. Int. Conf. Data Mining*, 2011, pp. 764–773.

[11] Y. Zheng, S. Neo, T. Chua, and Q. Tian, "Toward a higher-level visual representation for object-based image retrieval," *Vis. Comput.*, vol. 25, no. 1, pp. 13–23, 2009.

[12] S. Zhang *et al.*, "Modeling spatial and semantic cues for large-scale near-duplicated image retrieval," *Comput. Vis. Image Understand.*, vol. 115, no. 3, pp. 403–414, 2011.

[13] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3100–3107.

[14] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 1597–1604.

[15] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[16] B. Ni, M. Xu, J. Tang, S. Yan, and P. Moulin, "Omni-range spatial contexts for visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3514–3521.

[17] S. Singh, A. Gupta, and A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.

[18] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 9–16.

[19] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu, "3D visual phrases for landmark recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3594–3601.

[20] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2743–2750.

[21] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[22] Y. Lee and K. Grauman, "Object-graphs for context-aware visual category discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 346–358, Feb. 2012.

[23] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1745–1752.

[24] C. Li, D. Parikh, and T. Chen, "Automatic discovery of groups of objects for scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2735–2742.

[25] A. Gallagher and T. Chen, "Using context to recognize people in consumer images," *IPSJ Trans. Comput. Vis. Appl.*, vol. 1, pp. 115–126, Jan. 2009.

[26] A. Fitzgibbon and A. Zisserman, "On affine invariant clustering and automatic cast listing in movies," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 243–261.

[27] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jul. 2004, pp. 1–488.

[28] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Jan. 2010.

[29] G. Zhao, J. Yuan, J. Xu, and Y. Wu, "Discovery of the thematic object in commercial videos," *IEEE Multimedia Mag.*, vol. 18, no. 3, pp. 56–65, Mar. 2011.

[30] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool, "Efficient mining of frequent and distinctive feature configurations," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[31] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," *Data Mining Knowl. Discovery*, vol. 15, no. 1, pp. 55–86, 2007.

[32] S. Kim, X. Jin, and J. Han, "Disiclass: Discriminative frequent pattern-based image classification," in *Proc. 10th Int. Workshop Multimedia Data Mining*, 2010, pp. 1–7.

[33] B. Fernando, E. Fromont, and T. Tuytelaars, "Effective use of frequent itemset mining for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 214–227.

[34] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[35] A. Faktor and M. Irani, "'clustering by composition'-unsupervised discovery of image categories," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 474–487.

[36] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2009, pp. 2122–2129.

[37] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2003, pp. 264–271.

[38] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[39] B. Ommer and J. Buhmann, "Learning the compositional nature of visual objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[40] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2158–2174, Dec. 2008.

[41] M. Choi, A. Torralba, and A. Willsky, "A tree-based context model for object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 240–252, Feb. 2012.

[42] Y. Wu, Z. Si, H. Gong, and S. Zhu, "Learning active basis model for object detection and recognition," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 198–235, 2010.

[43] G. Bouchard and B. Triggs, "Hierarchical part-based visual object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 710–715.

[44] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[45] D. Parikh, C. Zitnick, and T. Chen, "Unsupervised learning of hierarchical spatial structures in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2743–2750.

[46] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[47] O. Boiman and M. Irani, "Similarity by composition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 177–184.

[48] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[49] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1800–1807.

[50] T. Malisiewicz and A. Efros, "Improving spatial support for objects via multiple segmentations," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2. 2007, pp. 1–3.

[51] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. Int. Conf. Image Video Retr.*, 2007, pp. 401–408.

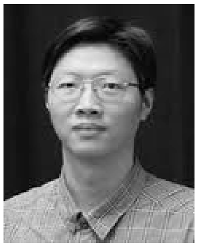
[52] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[53] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

Hongxing Wang (S'11) received the B.S. degree in information and computing science and the M.S. degree in operational research and cybernetics from Chongqing University, Chongqing, China, in 2007 and 2010, respectively.

He is currently pursuing the Ph.D. degree with Nanyang Technological University, Singapore. His current research interests include computer vision, image and video analysis, and pattern recognition.

Junsong Yuan (M'08) received the Ph.D. degree from Northwestern University, the M.Eng. degree from the National University of Singapore, and the B.Eng. degree from Special Class for the Gifted Young, Huazhong University of Science and Technology, China. He joins Nanyang Technological University as a Nanyang Assistant Professor in 2009, and currently leads the video analytics program with the School of EEE. His research interests include computer vision, pattern recognition, video analytics, multimedia search and data mining, and human–computer interaction. He has co-authored over 100 technical papers, and filed three U.S. patents and two provisional U.S. patents. He received the Outstanding EECS Ph.D. Thesis Award from Northwestern University and the Doctoral Spotlight Award from the IEEE Conference Computer Vision and Pattern Recognition Conference (CVPR'09). He is the Organizing Chair of the Asian Conference on Computer Vision (ACCV'14), Area Chair of WACV'14, ACCV'14, ICME'14, and Co-Chairs workshops at CVPR'12'13 and ICCV'13. He is an Associate Editor of the *Visual Computer Journal* and the *Journal of Multimedia*. He gives tutorials at the IEEE ICIP'13, FG'13, ICME'12, SIGGRAPH VRCAI'12, and PCM'12.



Ying Wu (SM'06) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 1994, 1997, and 2001, respectively.

From 1997 to 2001, he was a Research Assistant with the Beckman Institute for Advanced Science and Technology, UIUC. From 1999 to 2000, he was a Research Intern with Microsoft Research, Redmond, WA, USA. In 2001, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA, as an Assistant Professor. He was promoted as an Associate Professor in 2007 and a Full Professor in 2012. He is currently a Full Professor of Electrical Engineering and Computer Science with Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human–computer interaction.

He serves as an Associate Editors for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *SPIE Journal of Electronic Imaging*, and the *IAPR Journal of Machine Vision and Applications*. He received the Robert T. Chien Award by UIUC in 2001 and the NSF CAREER Award in 2003.