Omni-Range Spatial Contexts for Visual Classification

Bingbing Ni[1*], Mengdi Xu[2*], Jinhui Tang[3], Shuicheng Yan[2], Pierre Moulin[4]
[1]Advanced Digital Sciences Center, Singapore [2]National University of Singapore, Singapore
[3]Nanjing University of Science and Technology, China [4]University of Illinois at Urbana-Champaign, USA
bingbing.ni@adsc.com.sg {g0900224, eleyans}@nus.edu.sg
jinhuitang@mail.njust.edu.cn moulin@ifp.uiuc.edu

Abstract

Spatial contexts encode rich discriminative information for visual classification. However, as object shapes and scales vary significantly among images, spatial contexts with manually specified distance ranges are not guaranteed with optimality. In this work, we investigate how to automatically select discriminative and stable distance bin groups for modeling image spatial contexts to improve classification performance. We make two observations. First, the number of distance bins for context modeling can be arbitrarily large, and discriminative contexts are only from a small subset of distance bins. Second, adjacent distance bins for contexts modeling often show similar characteristics, thus encouraging grouping them together can result in more stable representation. Utilizing these two observations, we propose an omni-range spatial context mining framework for image classification. A sparse selection and grouping regularizer is employed along with an empirical risk, to discover discriminative and stable distance bin groups for context modeling. To facilitate efficient optimization, the objective function is approximated by a smooth convex function with theoretically guaranteed error bounds. The selected and grouped image spatial contexts, which are applied in food and national flag recognition, are demonstrated to be discriminative, compact and robust.

1. Introduction

Due to the compositional property of visual objects and scenes, mining discriminative spatial context patterns is of fundamental importance in visual classification. An appropriate combination of the stand-alone *weak* features can bring *strong* (*i.e.*, more discriminative) compositional features, as shown in [4, 5, 6, 10, 11, 12, 14, 16, 18, 23, 24, 25, 26, 27, 28].

To encode local spatial context, image local feature co-occurrences (*i.e.*, adjacent pixel pairs) are usually extracted for visual representation in image modeling and classification tasks [8, 10]. Li et al. [10] proposed a Markov stationary features method (MSF) which compactly encode co-occurrences of adjacent pixel pairs (using 4-pixel or 8-pixel neighborhood system) by computing stationary distributions. In addition to neighboring pixel pairs for calculating co-occurrence features, Ni et al. [16] defined a set of image local patterns (adjacent pixel set) for modeling higher order (*i.e.*, 3rd order) spatial contexts (CH). As in the popular bag-of-words model [9], images are represented as a set of sparsely located interesting points, *e.g.*, SIFT [13]. To encode spatial context for sparse features, in [25, 28], individual visual words (SIFTs) are grouped into visual phrases by searching each visual word's K-nearest neighbors. However, the co-occurrence representation is sensitive to the selection of the value K as shown in [25, 28].

Local spatial contexts (*e.g.*, adjacent pixels, nearest neighbors) often ignore richer information which can be described by feature pairs (set) that situate far from each other, *i.e.*, sometimes known as *far context*. For far context, not only the feature co-occurrence but also the geometric configuration of the context (*e.g.*, relative distance) give discriminative information. Spatial histogram is com-

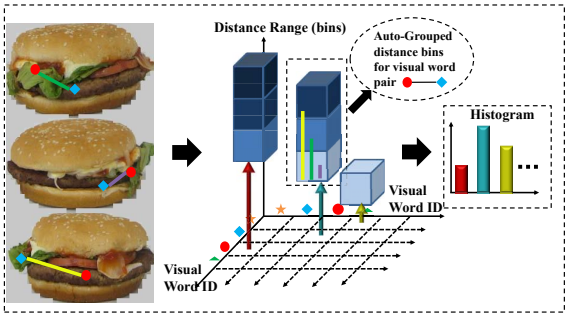

Figure 1. Illustration of the proposed omni-range spatial context mining framework. The distance bins for a specific visual word pair shall be automatically grouped by our proposed method and the sparsely selected distance bin groups are further used for ultimate visual classification. See color pdf for better view.

*indicates equal contribution.

monly used for spatial relationship computation between interest points [12, 26] situated with arbitrary distances. To calculate spatial histogram, local region centered on a visual feature is divided into several distance bins, *e.g.*, circular rings (linear or log scale or sometimes with directional bins). Zhang et al. [27] proposed a geometry-preserving visual phrases (GVP) method to capture the local and long-range spatial layouts of the visual words. Since objects may have various scales and shapes in different images, there is no guideline how to optimally select those context modeling parameters, *e.g.*, size and number of distance bins.

Ling et al. [11] proposed a Proximity Distribution Kernel (PDK), which embeds multiple range spatial co-occurrence information into a kernel representation. In [18], each visual feature is paired with a rectangular region with randomly generated size and displacement, and contextual information is encoded by responses of texture-layout filters on that rectangular region. In a more recent work of Shotton et al. [17], context information is represented by the depth relationship between depth pixel pairs which are located at a random displacement (*e.g.*, random direction and distance). These methods explore different distance ranges for spatial context modeling, however, the resulting representations are highly complex and redundant. In other works [4, 5, 6], spatial context is represented by a joint Gaussian density of the locations of features within a hypothesis. In these models, absolute image feature locations (instead of pair-wise relationship) are used, which are inflexible as object positions, orientations and scales have large variations in general images.

We have the following observations. First, values of parameters in spatial context modeling such as the number of distance bins and the bin size can be arbitrary. However, often there are only a small number of configurations (sparse) which can give optimal performance. It is therefore demanding to have a selection method which can efficiently capture those highly discriminative configurations for context modeling. In contrast, previous methods either fix a certain configuration for calculating context [10, 12, 16, 25, 26, 28] or use highly redundant representations by enumerating all distance ranges [11, 17, 18] (TextonBoost [18] has a post-processing step of context selection for boosting classification performance). Second, to cope with the object scale variations, important geometric prior knowledge can be utilized. More specifically, spatial contexts from adjacent distance bins usually share similar characteristics and grouping them can not only improve stability, robustness in context modeling (for avoiding overfitting caused by over-fine binnings), but also provide more compact context representation. Previous methods, however, do not explore such a grouping mechanism. Table 1 compares state-of-the-art context modeling methods with our proposed method in terms of 1,2) whether it is capable of local or far context modeling? 3)whether it is capable of encoding high-order contexts? and 4) whether the distance ranges (or k-nearest-neighbors) for context modeling are manually fixed or automatically learned.

A recently proposed ℓ_1/ℓ_∞-regularizer (OSCAR) [1] sheds some lights on addressing these problems. The OSCAR regularizer has the capability of doing sparse and grouped variable selection by encouraging correlated predictors that have a similar effect on the response to form predictive clusters represented by a single coefficient. In the same spirit, only a small subset of the whole context collection convey discriminative information, and geometrically adjacent distance partitions (bins) of contexts often have similar effects (contributions) on the discriminative power; therefore, the same ℓ_1/ℓ_∞-regularizer could be applied for encouraging merging them. The resulting context representation will be much more compact and more robust to image and object variations. Apart from the general OSCAR formulation where there exist no prior information on the correlations among the variables to group and select, in our problem, strong correlations exist between adjacent distance bins of spatial contexts. Therefore, preserving this important geometric property favors a more suitable solution. To utilize this spatial smoothness prior information, we only allow pair-wise ℓ_∞ regularization defined for geometrically adjacent context bins, *i.e.*, adjacent distance bins of context. Exploiting these properties, we propose an omni-range spatial context mining framework. We first partition the context configuration space in a fine resolution, *i.e.*, with fine distance binnings. We then unify the context mining framework by employing an ℓ_1/ℓ_∞-regularized squared hinge loss objective, with geometric structure prior. To efficiently optimize this cost function, a smooth approximation to the ℓ_1/ℓ_∞-regularizer with theoretically guaranteed error bounds is introduced and solved by accelerated proximal gradient method [19]. The proposed omni-range context modeling framework is extensively applied on two image classification tasks including food and national flag recognition. Its discriminative capability, robustness and compactness are demonstrated. Figure 1 illustrates the proposed method.

2. Omni-Range Spatial Context Mining

2.1. Problem Formulation

We consider the following setup in the context of binary image classification and object recognition. Assume there are N images. For a given image or object class label, images belonging to that class or contain the object are denoted as positive samples and the rest are regarded as negative samples. We denote image i as I_i where the label $l(I_i) = y_i \in \{+1, -1\}$. We extract a set of visual features $\mathbf{p}_i$ (*e.g.*, SIFT) from each image by either interest

Table 1. Comparisons of different context modeling methods.

Method	VP [25]	MSF [10]	CH [16]	PDK [11]	GPVP [27]	TB [18]	Ours
Local Context?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Far Context?	No	No	No	Yes	Yes	Yes	Yes
High-Order Context?	Yes	No	Yes	No	Yes	No	No
Auto-Range?	No	No	No	No	No	Yes	Yes

point detection or sampling on regular grid; each feature point $\mathbf{p}_i$ is represented by its descriptor and its $2D$ position. As in the bag-of-words model, all features are vector quantized into a dictionary of size K, *i.e.*, K visual word prototypes $vw_1, vw_2, \cdots, vw_K$; each feature i is then indexed by searching the nearest visual word prototype.

Centered at each feature on the image, we mask a circle-type spatial histogram with distance bined in linear scale, as illustrated in Figure 2 (note that directional bin can also be used). In Figure 2, context is described by the co-occurrence of two visual words (vw_i, vw_j) with their related distance bin index r, where $r = 1, 2, \cdots, R$ (R is the number of distance bins). Note that fine partition for context configuration space is used since our consecutive context mining framework is capable of merging similar-effect distance bins to achieve more compact representation. Each image could be initially represented as a histogram vector $\mathbf{x}$ where each element corresponds to the normalized frequency on a context triplet (vw_i, vw_j, r), *i.e.*, the length of the image representation is $D = K \times K \times R$.

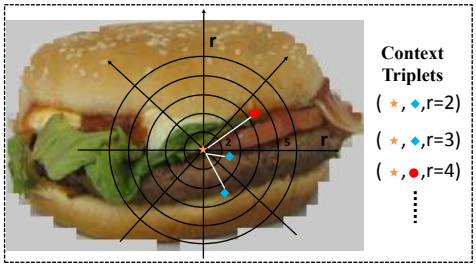

Figure 2. Diagrammatic illustration of spatial context (co-occurrence) triplet representation.

We embed the omni-range context mining objective within a learning framework. The ultimate learning goal is to obtain a classification function $f(\mathbf{x})$, where $f(\mathbf{x}) > 0$ indicates positive samples and $f(\mathbf{x}) < 0$ indicates negative samples. An common choice of $f(\mathbf{x})$ is linear classifier given by: $f(\mathbf{x}|\mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ where $\mathbf{w} \in \mathbb{R}^D$ is the desired hyperplane parameter vector, b is the bias term and $\langle . \rangle$ denotes dot product. A cost function E will be defined based on f for the objective of distinguishing two classes. The obtained $\mathbf{w} \in \mathbb{R}^D$ could be also considered as the responses on the input elements (contexts), where strong response indicates more relevant in distinguishing two classes, thus need to be retained. As will be described in detail as follows, how to regularize the cost function E lies in the heart of the problem formulation.

2.2. Obj-I: Empirical Risk for Classification

The ultimate goal of mining spatial context is to maximize the discriminative capability for the selected contexts. We therefore consider the following empirical risk directly targeting at boosting classification performance for f:

$$E(\mathbf{w}) := \sum_{i=1}^{N} V(y_i, \langle \mathbf{w}, \mathbf{x}_i \rangle + b), \tag{1}$$

where $V(u, v) := \frac{1}{2}[1 - uv]_+^2$ is the square of hinge loss with $[\cdot]_+$ denoting the operation of $\max\{0, \cdot\}$. N is the number of images. $E(\mathbf{w})$ is convex, continuous and first-order differentiable. Its gradient can be given by:

$$\frac{\partial E(\mathbf{w}, b)}{\partial \mathbf{w}} : = -\sum_{i=1}^{N} y_i(1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))\mathbf{x}_i, \tag{2}$$

$$\frac{\partial E(\mathbf{w}, b)}{\partial b} : = -\sum_{i=1}^{N} y_i(1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)). \tag{3}$$

Its Lipschitz constant is given by $\|X^T X\|_F$ ($\|.\|_F$ denotes the Frobenius norm), where $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$.

2.3. Obj-II: Sparse Selection and Grouping Regularization

We have two key observations: 1) although the dimensionality of the spatial context configuration space (*i.e.*, number of distance bins) could go into extremely large, however, only a small subset of the whole context collection conveys discriminative information, therefore the result of feature (context) selection is expected to be sparse; and 2) as there exist significant variations in scale, orientation in images/objects, adjacent distance bins in context modeling often show similar effects (contribution) on the discriminative power; therefore, it is practically sound to merge context triplets, *i.e.*, (vw_i, vw_j, r)s into stable groups according to the adjacency of distance bins (r).

The proposed regularization framework addresses these problems. An ℓ_1-regularizer which is applied on all variables (contexts) is combined with pairwise ℓ_∞-regularizers which are defined on spatially adjacent distance bins in context triplets. For example, (vw_i, vw_j, r) and $(vw_i, vw_j, r+1)$ could be potentially merged if they receive similar coefficients. Sharing similar principle with the OSCAR regularizer [1], the proposed regularization framework jointly encourages sparsity as well as grouping of adjacent contexts which show similar characteristics. Formally, this proposed

regularization term $R(\mathbf{w})$ can be formulated as:

$$R(\mathbf{w}) = \lambda_1 R_1(\mathbf{w}) + \lambda_2 R_2(\mathbf{w}), \quad (4)$$

$$= \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \sum_{i,j \in \mathcal{N}} \max\{|w_i|, |w_j|\}, \quad (5)$$

where $\mathcal{N}$ denotes the set of adjacent context bins. It is trivial to verify that $R(\mathbf{w})$ is convex and non-differentiable. λ_1 and λ_2 are the corresponding weighting factors, and in this work they are optimally set via cross validation.

2.4. A Regularized Omni-Range Context Mining Framework

We are now in the position to formulate a unified framework for regularized sparse context selection and grouping by composing (1) and (4):

$$F(\mathbf{w}) = E(\mathbf{w}) + R(\mathbf{w}), \quad (6)$$

$$= \sum_{i=1}^{N} V(y_i, \langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$

$$+ \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \sum_{i,j \in \mathcal{N}} \max\{|w_i|, |w_j|\}. \quad (7)$$

It is straightforward to verify that the objective $F(\mathbf{w})$ is convex but non-smooth since the regularizers are non-smooth. The original OSCAR formulation could be re-formulated as a quadratic programming problem with $O(D^2)$ linear constraints (*i.e.*, D is the total number of variables). Although sparse, the constraint matrix is very large, using the off-the-shelf quadratic programming algorithm SQOPT [7] can not efficiently solve the problem. By exploring the symmetric property of the ℓ_1/ℓ_∞ regularizer, Zhong and Kwok [29] proposed an efficient solver based on the accelerated gradient methods [19] where the key projection step can be solved by a simple iterative group merging algorithm. For our problem, since symmetry no longer holds due to the introduction of spatial prior, this efficient iterative group merging algorithm is not applicable. In this work, we utilize recent advances in smoothing approximation method [15] to relax our formulation to a convex and smooth one with theoretical guarantee on the error bounds, where efficient optimization is attainable, as will be presented in the next subsection.

2.5. Smoothness Approximation and Optimization

Based on the smoothing approximation techniques originally from [15], the ℓ_∞-regularizer term $\|\mathbf{u}\|_\infty$ can be approximated by the following smooth function:

$$q_\mu(\mathbf{u}) = \max_{\|\mathbf{v}\|_1 \leq 1} \{\langle \mathbf{u}, \mathbf{v} \rangle - \frac{1}{2}\mu \|\mathbf{v}\|_2^2\}. \quad (8)$$

Herein, μ is a parameter to control the approximation accuracy. The following proposition justifies the max structure equivalence of ℓ_∞-norm.

Proposition 1 *For any vector $\mathbf{u} \in \mathbb{R}^d$, there is a constant vector $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{u}\|_\infty$ has a max-structure representation in the following form,*

$$\|\mathbf{u}\|_\infty = \max_{0 \leq \|\mathbf{v}\|_1 \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle,$$

where $\|\mathbf{v}\|_1$ denotes ℓ_1-norm of $\mathbf{v}$.

Proof 1 *Without loss of generality, we define $u_m = \max_i |u_i|$. By definition we have that:*

$$\|\mathbf{u}\|_\infty = \max_i |u_i| = u_m = \max_{0 \leq \|\mathbf{v}\|_1 \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle, \quad (9)$$

where the last equality follows the fact that to maximize $\langle \mathbf{u}, \mathbf{v} \rangle$ under $0 \leq \|\mathbf{v}\|_1 \leq 1$ constraint, only the element with the maximum absolute value in $\mathbf{u}$ is selected.

For a fixed $\mathbf{u}$, we denote $\mathbf{v}(\mathbf{u})$ as the unique optimizer of (8). Then $\mathbf{v}(\mathbf{u})$ can be very easily obtained via the ℓ_1-ball projection algorithm as stated in [3]. For the two variable case ($d = 2$, $\mathbf{u} = (u_1, u_2)$, which is our case, *i.e.*, pair-wise) $\mathbf{v}(\mathbf{u})$ can be obtained analytically as:

$$\mathbf{v}(\mathbf{u}) = \begin{cases} (\frac{u_1}{\mu}, \frac{u_2}{\mu}), |u_1| + |u_2| \leq \mu; \\ (sign(u_1)(\frac{1}{2} + \frac{|u_1|-|u_2|}{2\mu}), sign(u_2)(\frac{1}{2} + \frac{|u_2|-|u_1|}{2\mu})), \\ |u_1| + |u_2| > \mu, -\mu \leq |u_1| - |u_2| \leq \mu; \\ (sign(u_1), 0), |u_1| - |u_2| > \mu; \\ (0, sign(u_2)), |u_1| - |u_2| < -\mu. \end{cases}$$

The four cases are illustrated in Figure 3.

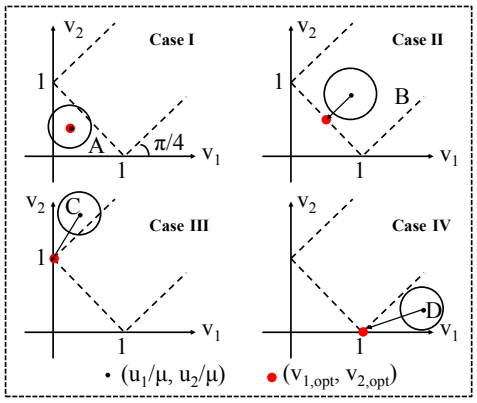

Figure 3. Four cases for the optimal solution of $\mathbf{v}(\mathbf{u})$, in the scenario of $d = 2$.

The following proposition gives bounds to the approximation.

Proposition 2 *$q_\mu(\mathbf{u})$ is a μ-accurate approximation to $q(\mathbf{u}) = \|u\|_\infty = \max_{0 \leq \|\mathbf{v}\|_1 \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle$, that is*

$$q_\mu(\mathbf{u}) \leq q(\mathbf{u}) \leq q_\mu(\mathbf{u}) + \frac{1}{2}\mu. \quad (10)$$

Moreover $R_2(\mathbf{w}) = \sum_{i,j\in\mathcal{N}} \max\{|w_i|, |w_j|\}$ *is bounded by:*

$$\sum_{i,j\in\mathcal{N}} q_\mu((w_i, w_j)) \leq R_2(\mathbf{w}) \leq \sum_{i,j\in\mathcal{N}} q_\mu((w_i, w_j)) + \frac{\mu M}{2}, \tag{11}$$

where M *is the number of pair-wise regularization terms.*

Proof 2 *Since* $0 \in \{v_j : 0 \leq v_j \leq 1\}$*, by definition we get that*

$$0 \leq q_\mu(\mathbf{u}) \leq \max_{0\leq\|\mathbf{v}\|_1\leq 1} \langle \mathbf{v}, \mathbf{u}\rangle \tag{12}$$

we can also have

$$q_\mu(\mathbf{u}) \geq \max_{0\leq\|\mathbf{v}\|_1\leq 1} \langle \mathbf{v}, \mathbf{u}\rangle - \frac{1}{2}\mu. \tag{13}$$

Combining (12) *and* (13) *we get*

$$q_\mu(\mathbf{u}) \leq \max_{0\leq\|\mathbf{v}\|_1\leq 1} \langle \mathbf{v}, \mathbf{u}\rangle = q(\mathbf{u}) \leq q_\mu(\mathbf{u}) + \frac{1}{2}\mu. \tag{14}$$

Summing up (14) *over all pair-wise regularization terms in* $R_2(\mathbf{w})$ *gives* (11).

Moreover, from the standard results of [Theorem 1] in [15]) $q_\mu(\mathbf{u})$ is differentiable and its gradient $q'_\mu(\mathbf{u}) = \mathbf{v}(\mathbf{u})$ is Lipschitz continuous with the constant $\frac{1}{\mu}$. If we denote $G_\mu(\mathbf{w}) = \sum_{i,j\in\mathcal{N}} q_\mu((w_i, w_j))$, which is the μ-approximation to $R_2(\mathbf{w})$, we can have its gradient $\nabla G_\mu(\mathbf{w}) = \sum_{i,j\in\mathcal{N}} \mathbf{v}(w_i, w_j)$, with Lipschitz constant as $L_{G_\mu} = \frac{M}{\mu}$ by triangular inequality.

Similarly, for the ℓ_1 component of $R(\mathbf{w})$ (*i.e.*, $R_1(\mathbf{w})$), according to the same standard approximation result given in [15], we can approximate it by

$$P_\mu(\mathbf{w}) = \max_{0\leq\|\mathbf{v}\|_\infty\leq 1} \{\langle \mathbf{w}, \mathbf{v}\rangle - \frac{1}{2}\mu\|\mathbf{v}\|_2^2\}, \tag{15}$$

where its gradient can be denoted as $\nabla P_\mu(\mathbf{w}) = \mathbf{v}(\mathbf{w})$ ($\mathbf{v}(\mathbf{w})$ is the optimizer of (15)) with Lipschitz constant $L_{P_\mu} = \frac{D}{\mu}$. It is straightforward to see $P_\mu(\mathbf{w})$ is bounded as $\|\mathbf{w}\|_1 \leq P_\mu(\mathbf{w}) \leq \|\mathbf{w}\|_1 + \frac{\mu D}{2}$. Therefore the objective function $F(\mathbf{w})$ can be approximated by

$$F_\mu(\mathbf{w}) = E(\mathbf{w}) + \lambda_1 P_\mu(\mathbf{w}) + \lambda_2 G_\mu(\mathbf{w}), \tag{16}$$

with gradient

$$\nabla F_\mu(\mathbf{w}) = \nabla E(\mathbf{w}) + \lambda_1 \nabla P_\mu(\mathbf{w}) + \lambda_2 \nabla G_\mu(\mathbf{w}). \tag{17}$$

The overall Lipschitz constant L_{F_μ} can be therefore derived as: $L_{F_\mu} = \|X^T X\|_F + \lambda_1 \frac{D}{\mu} + \lambda_2 \frac{M}{\mu}$. To optimize the smoothed objective, we can then apply the accelerated proximal gradient method (APG), which has the convergence rate of $O(\frac{L_{F_\mu}}{\epsilon})$ [19] in terms of the desired residues, *i.e.*, $|F_\mu - \min F_\mu| \leq \epsilon$. By choosing $\mu \approx \epsilon$, we have that the rate of convergence is $O(1/\epsilon)$. The APG-based optimization procedure is given in Algorithm 1.

Inputs : $X \in \mathbb{R}^{D\times N}$,$\lambda_1$, λ_2, μ, $\{y_i, i = 1, \cdots, N\}$.
Output: $\mathbf{w}$.
Initialization: Calculate $L_{F_\mu} = L_f + \lambda_1 L_{G_\mu} + \lambda_2 L_{P_\mu}$.
Initialize $\mathbf{w}_0, \boldsymbol{\beta}_0 \in \mathbb{R}^D$, and let $\gamma_0 \leftarrow 0$, $k \leftarrow 0$.
repeat
  $\boldsymbol{u}_k = (1-\gamma_k)\mathbf{w}_k + \gamma_k \boldsymbol{\beta}_k$,
  Calculate the gradient of the smoothed function $F_\mu(\boldsymbol{u}_k)$, denoted as $\nabla F_\mu(\boldsymbol{u}_k)$.
  $\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{1}{\gamma_k L_{F_\mu}} \nabla F_\mu(\boldsymbol{u}_k)$,
  $\mathbf{w}_{k+1} = (1-\gamma_k)\mathbf{w}_k + \gamma_k \boldsymbol{\beta}_{k+1}$,
  $\gamma_{k+1} = \frac{2}{k+1}$, $k \leftarrow k+1$.
until *Converged*;

Algorithm 1: Optimization procedure for objective (16).

2.6. Application to Image Classification

The obtained coefficients $\mathbf{w}$ is sparse and present group patterns. First we eliminate the contexts corresponding to the zero elements of $\mathbf{w}$. Then, starting from single element groups, we progressively merge groups of contexts by using the information from: 1) values of non-zero elements of $\mathbf{w}$, and 2) the geometrical adjacency relationship of distance bins of contexts. The output are the groups of context triplets which can be used for image/object representations in classification. More specifically, we denote $\mathbf{w}' \in \mathcal{R}^{D'}$ as the non-zero elements of the learned $\mathbf{w}$, and $G_1, G_2, \cdots, G_{D'}$ as the initial context groups where each G_i only contains i-th element of $\mathbf{w}'$. Two groups G_i and G_j are merged according to the condition:

$$\|\hat{w}(G_i) - \hat{w}(G_j)\|_2 \leq \delta, \tag{18}$$

$$G_i \in \mathcal{N}(G_j), \tag{19}$$

where δ is a threshold value (which is set empirically). Here $\hat{w}(G_i)$ denotes the mean coefficient value of the elements in G_i as

$$\hat{w}(G_i) = \frac{1}{|G_i|} \sum_{w_j\in G_i} w_j, \tag{20}$$

where $|G_i|$ denotes the number of elements in G_i. $G_i \in \mathcal{N}(G_j)$ means that as least one element of both groups are adjacent. The adjacency condition ensures that only adjacent distance bins can be merged. The merging procedure is performed progressively until there exist no two groups that can be merged. Each obtained context group correspond to a list of context triplet (vw_i, vw_j, r). For representing an image, frequencies (counted numbers) of the context triplets within the same group are summed up, and the resulting representation is a D_G-dimensional vector (histogram), where $D_G = |\{G_i\}|$ is the number of groups after the merging procedure. The derived context groups can significantly reduce the complexity of context representation, as the number of retained groups can be much

smaller than the number of (vw_i, vw_j, r) triplets. It also achieves more stable representation under the scenario of object shape/scale variation owing to its geometric-aware grouping property. These derived context groups are called *Omni-Range Contexts* in this work.

There exist a lot of interesting visual classification methods, such as [20, 21]. Since image representation is the main focus of our work, here we adopt a simple multi-classification approach as follows: we first derive omni-range contexts for each category, and then we train a binary support vector machine classifier based on χ^2 distance kernel for each class (negative samples are randomly selected from other categories). Suppose we have C classes, the SVM output scores of the C classes are then used as a middle-level feature representation to train a multi-class support vector machine classifier using RBF kernel. For training SVMs, the penalty parameter C is optimized by cross-validation and the bandwidth parameters for χ^2 and RBF kernels are set as the average of the squared χ^2 or Euclidean distances of the training sample pairs.

3. Experiments

In this section, we present our experiments on two image classification tasks including food and national flag recognition. The reason to apply the proposed omni-range context framework on these two problems is because: 1) a food item can largely be characterized by its ingredients and their relative spatial relationships, as indicated in [22]; and 2) the visual elements across different national flags are similar, however the relative spatial relationships are distinctive.

3.1. Food Recognition

For food recognition, we use the Pittsburgh Food Image Dataset (PFID) dataset [2]. The PFID dataset is a collection of fast food images and videos from 13 chain restaurants acquired under lab and realistic settings. Same as in [2, 22], our experiments focus on the set of 61 categories of specific food items (e.g., McDonalds Big Mac) with masked background. Each food category contains three different instances of the food (bought on different days from different branches of the restaurant chain), and six images from six viewpoints (60 degrees apart) of each food instance. We follow the same experimental protocol used in [2, 22] and perform 3-fold cross-validation for our experiments, using the 12 images from two instances for training and the 6 images from the third for testing. This procedure is repeated three times, with a different instance serving as the test set and the results are averaged. SIFT descriptors are uniformly sampled from images by a step of 4 pixels. Feature descriptors are collected and vector-quantized using Kmeans clustering. The resulting cluster centers form the dictionary of visual words. For the dictionary size, we use $K = 50$, as our experiments show higher values of K do not bring significant performance improvement. Each local feature descriptor is then indexed by its closest codeword. Note that we use dense SIFT features instead of sparsely detected features as the latter give significantly worse performances for all comparing algorithms in our experiments. For the initial context configuration, we use the spatial histogram in Figure 2. We used 16 distance bins (as a good compromise between efficiency and accuracy) with equal distance spacing of 5 pixels. For training omni-context model for each category, we randomly selected negative images from the rest classes in the dataset, and we keep the ratio of number of positive samples vs. number of negative samples as $(1 : 3)$. We evaluate the performances in terms of classification accuracy.

We implement the optimization procedure using Matlab 2010 on a 2.63 Ghz machine with 8GB of memory. Figure 4 shows one typical convergence curve of Algorithm 1 on one category of the dataset where the X-axis indicates the increased iteration steps and the Y-axis indicates the objective function values. The procedure converges typically within $10,000$ iterations, which takes about 25 seconds on our computing platform.

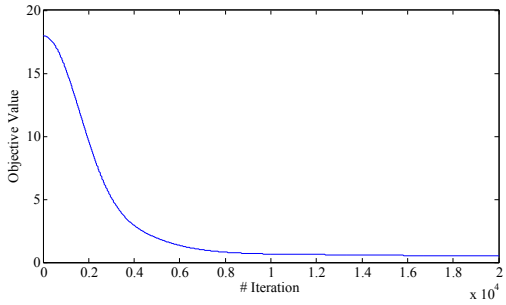

Figure 4. A typical convergence curve of our optimization procedure on the food dataset.

We visualize trained coefficients for high frequency visual word pairs (2500 visual word pairs) in Figure 5. From Figure 5, we can observe that the selected features are sparse and for most visual word pairs, contexts from adjacent distance bins receive similar coefficients, which can be grouped by our merging procedure.

We compare the classification accuracies achieved by the proposed omni-range context framework with the state-of-the-art result in [22] (denoted as orientation and midpoint category pairwise features (OM)). We also implement and compare the following commonly used baseline and spatial context modeling approaches: 1) the baseline BOW representation; 2) spatial pyramid matching kernel (SPMK) [9]; 3) Markov stationary features (MSF) [10]; 4) contextualized histogram (CH) [16]; 5) proximity distribution kernel (PDK) [11]; 6) TextonBoost (TB) [18]; and 7) geometry-preserving visual phrases (GPVP) [27]. Cross-validation has been performed to determine the free parameters for each algorithm. Table 2 illustrates classification accuracies by different algorithms. We can see that owing to its ca-

Table 2. Classification accuracies and the corresponding representation complexities using different algorithms on food recognition. For codebook size K, the dimensionality is K, $21K$ (3 levels), $2K$, $30K$, $K^2 \times R$ for BOW, SPMK, MSF, CH and PDK, respectively. For TB, the complexity corresponds to the number of random pair-wise contexts. For our method, the complexity means the number of retained context groups.

Algorithm	OM	BOW	SPMK	MSF	CH	PDK	TB	GPVP	Ours
Accuracy (%)	28.2	30.6	33.4	32.7	34.9	35.3	38.4	35.9	**42.9**
Dimensionality	6144	50	1050	100	1500	40000	54732	100K	15452

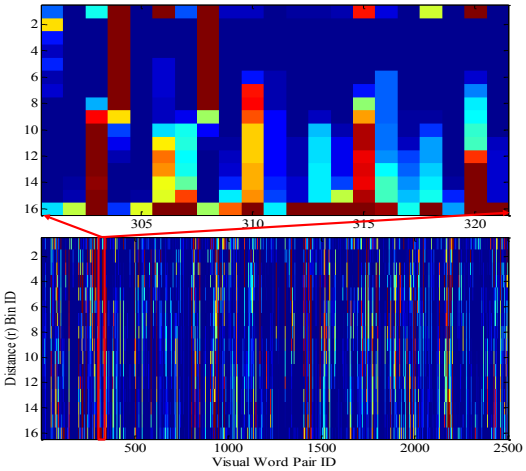

Figure 5. Visualization of the learned coefficients for visual word triplets (vw_i, wv_j, r)s with zoom-in regions. See color pdf for better view.

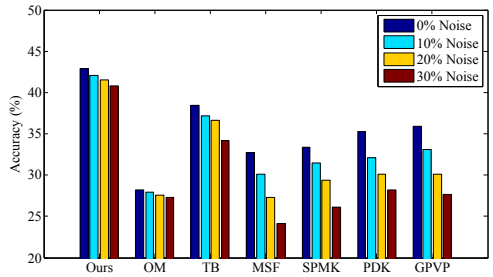

Figure 6. Classification accuracies for different algorithms on different levels of random scalings for the testing images.

Figure 7. Sample images of the national flag dataset. Note there exist significant scale/orientation variations and deformations over different images.

pability of context selection and grouping, our algorithm achieves the highest accuracy. In Table 2 the corresponding representation complexities of different methods (in terms of feature dimensionality) are also given. One can see that although PDK, TB and GPVP achieve high classification accuracies, their representation complexities are high. In contrast, our proposed method has a good trade-off between accuracy and complexity. Also, it is interesting to note that simple BOW based method can outperform the state-of-the-art result by the OM method. To evaluate the robustness, we randomly scale the testing images by zero mean Gaussian distributed delta scale values, the variances of the delta scale values are set as 10%, 20%, 30% of the original image scales. We then apply the previously trained omni-range context models on the distorted testing images. We compare different algorithms and the results are illustrated in Figure 6. We can see that our algorithm scales well with random scaling owing to its capability in grouping adjacent distances bins for the sake of stability. OM and TB also present scale invariance properties. Other methods such as MSF, SPMK and GPVP are sensitive to scale variations.

3.2. National Flag Classification

We collect a national flag dataset which contains totally 8298 national flag images belong to 39 countries. These images are crawled from Google images using text queries, and the image size in average is about 350×200 pixels. Figure 7 shows several sample images from the dataset and we can notice that there exist large variations on the sizes (scales) and orientations of the flags and in most cases, the flags are deformed, which further challenges the visual classification performance. The dataset is randomly partitioned into equally-sized training and testing set. For training images, we manually crop the flags, mark 4 corners and 4 midpoints. We use these landmarks to normalize the flags. The normalization is performed by transforming (translation, rotation, scaling) all cropped flags onto a standard size using the corner points' positions. For all images, we use Lowe's SIFT detector [13] to detect interesting points and extract SIFT descriptors. Same as in the food recognition experiment, we index the features using the bag-of-words method with dictionary size $K = 50$ (which is empirically optimal). For testing, we use a sliding window which has different scales and searches the entire image with step size of 5 pixels. When the sub-image in the searching window gives a classification score higher than the trained threshold, it is decided as detection. We require the detected window overlaps more than 75% with the ground truth window in calculating accuracy. As in the food recognition experiment, we compare the recognition accuracies (along with the cor-

Table 3. Classification accuracies and the corresponding representation complexities using different algorithms on national flag recognition.

Algorithm	BOW	SPMK	MSF	CH	PDK	TB	GPVP	Ours
Accuracy (%)	32.5	35.7	34.8	37.9	36.9	45.6	39.2	**53.4**
Dimensionality	50	1050	100	1500	40000	67232	100K	13984

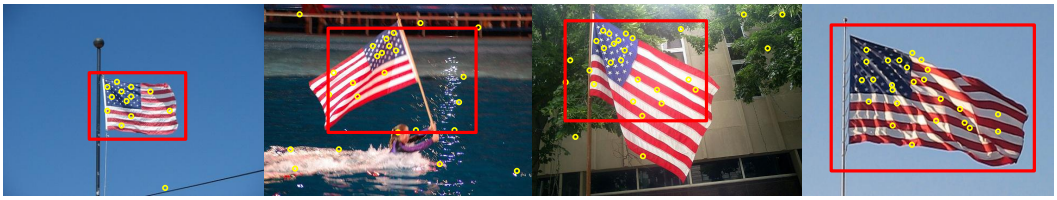

Figure 8. Examples of national flag recognition results. The red rectangles show under which scales our method successfully recognizes the flags. Yellow dots are the detected SIFT points.

responding representation complexities) of different algorithms, which are shown in Table 3. We can observe that the proposed omni-range context method outperforms other state-of-the-art methods. Figure 8 shows several examples of recognition results using our method, where we can observe that since our method encodes distance (range) information for contexts, the correct target object can automatically *pop out* when the searching window's scale matches the trained omni-context model. In this sense, the proposed method can be considered as having a byproduct of object detection.

4. Conclusions

In this work, we proposed an omni-range context mining framework for automatically selecting and grouping distance bins in spatial context modeling. The derived omni-range contexts, which are extensively applied on two visual classification tasks including food recognition and national flag recognition, have demonstrated to be discriminative, robust and compact.

Acknowledgement

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR). It is also supported by NSFC under grant 61103059.

References

[1] H. B. abd B.J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008. 2, 3

[2] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. Pfid:pittsburgh fast-food image dataset. In *ICIP*, 2009. 6

[3] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *ICML*, 2008. 4

[4] L. Fei-fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004. 1, 2

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1, 2

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 1, 2

[7] P. E. Gill, W. Murray, and M. A. Saunders. Users guide for sqopt 7: A fortran package for large-scale linear and quadratic programming. *Technical Report NA 05-1, Department of Mathematics, UCSD*, 2005. 4

[8] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *T-SMC*, 3(6):610–621, 1973. 1

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 6

[10] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histogram: Image representation using markov stationary features. In *CVPR*, 2008. 1, 2, 3, 6

[11] H. Ling and S. Soatto. Proximity distribution kernels for geometric context in category recognition. In *ICCV*, 2007. 1, 2, 3, 6

[12] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008. 1, 2

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, 2004. 1, 7

[14] T. Mita, T. Kaneko, B. Stenger, and O. Hori. Discriminative feature co-occurrence selection for object detection. *T-PAMI*, 30(7):1257–1269, 2008. 1

[15] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, (103):127–152, 2005. 4, 5

[16] B. Ni, S. Yan, and A. Kassim. Contextualizing histogram. In *CVPR*, 2009. 1, 2, 3, 6

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011. 2

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81:2–23, 2009. 1, 2, 3, 6

[19] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM J. Optim.*, 2008. 2, 4, 5

[20] M. Wang, X.-S. Hua, R. Hong, J. Tang, and Y. S. Guo-Jun Qi. Unified video annotation via multi-graph learning. *T-CSVT*, 19(5):733–746, 2009. 6

[21] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *T-MM*, 11(3):465–476, 2009. 6

[22] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *CVPR*, 2010. 6

[23] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *ICCV*, 2011. 1

[24] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1

[25] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007. 1, 2, 3

[26] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia*, pages 75–84, 2009. 1, 2

[27] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, 2011. 1, 2, 3, 6

[28] Y. Zheng, M. Zhao, S. Neo, T. Chua, and Q. Tian. Visual synset: towards a higher-level visual representation. In *CVPR*, 2008. 1, 2

[29] L. W. Zhong and J. T. Kwok. Efficient sparse modeling with automatic feature grouping. In *ICML*, 2011. 4