Boosted random contextual semantic space based representation for visual recognition

Chunjie Zhang¹, Zhe Xue¹, Xiaobin Zhu², Huanian Wang³, Qingming Huang³, Qi Tian⁴

¹School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
²Beijing Technology and Business University, Beijing, China
³Central University of Finance and Economics, Changping District, Beijing, China
⁴Key Lab of Intelligent Information Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
⁵Department of Computer Sciences, University of Texas at San Antonio, TX, 78249, USA

A R T I C L E   I N F O

Article history:
Received 16 October 2015
Revised 12 May 2016
Accepted 19 June 2016
Available online 21 June 2016

Keywords:
Pattern recognition
Image processing
Visual representation

A B S T R A C T

Visual information has been widely used for image representation. Although proven very effective, the visual representation lacks explicit semantics. However, how to generate a proper semantic space for image representation is still an open problem that needs to be solved. To jointly model the visual and semantic representations of images, we propose a boosted random contextual semantic space based image representation method. Images are initially represented using local feature’s distribution histograms. The semantic space is generated by randomly selecting training images. Images are then mapped into the semantic space accordingly. Semantic context is explored to model the correlations of different semantics which is then used for classification. The classification results are used to re-weight training images in a boosted way. The re-weighted images are used to construct new semantic space for classification. In this way, we are able to jointly consider the visual and semantic information of images. Image classification experiments on several public datasets show the effectiveness of the proposed method.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Visual information [35] has been widely adopted for various tasks, e.g. classification [35], segmentation [22] and retrieval [33]. Local feature based strategy is widely used for its efficiency. Usually, local features are encoded with pre-learned codebooks. A number of local features (e.g. SIFT [25], HoG [6], SURF [2], MROGH [10] and KAZE [1]) have been proposed whose effectiveness has been proven by researchers.

Although effective, visual feature has no explicit semantic correspondence with human perception. Many methods [7,45,56] only use visual cues for representations. However, due to the semantic gap [37], only using visual information cannot model semantics well. Hence, how to explore the semantic information of images becomes urgent. The state-of-the-art semantic based methods try to solve this problem with training images [4,14,20,23,28,30,43,45,52,55] or using images

* Corresponding author.

E-mail addresses: zhangcj@ucas.ac.cn (C. Zhang), xuezhe10@mails.ucas.ac.cn (Z. Xue), brucezhucas@gmail.com (X. Zhu), huanianwang06@gmail.com (H. Wang), qmhuang@jdl.ac.cn (Q. Huang), qitian@cs.utsa.edu (Q. Tian).

http://dx.doi.org/10.1016/j.ins.2016.06.029
0020-0255/© 2016 Elsevier Inc. All rights reserved.
from other sources [21,31,39,49,51]. Using the training images can help to generate semantic spaces discriminatively. However, most of them only consider the visual information for semantic space construction. The initially generated semantic space may not be able to represent images well. Besides, it is hard to learn effective classifiers for generic image classes. Moreover, object often exhibits visual polysemy. The leverage of extra information [21,31,39,49,51] can make use of more information along with the training images. However, this approach is still dataset dependent once the images are collected. Besides, the collected images may be improper for the specific task.

To alleviate the semantic gap, attribute based image representation also becomes popular [15–17,36]. Attributes are concepts that can be understood by human being and are also easy to be distinguished by computers. This helps to represent images in an understandable way. However, attributes have to be pre-defined by experts. Besides, there are many concepts that cannot be well represented by pre-defined attributes.

To solve the problems mentioned above, in this paper, we propose a novel boosted random contextual semantic space based image representation method for visual classification (BRCSS). Images are initially represented with the bag-of-visual-words (BoW) model. The semantic space is then generated by random selection. Contextual semantic representations of images are used to model the correlations of different semantics. We then train classifiers for prediction. We use the results in a boosted way by re-weighting images which are then used for semantic space construction. We conduct classification experiments on several public available image datasets, experimental results prove the effectiveness of the proposed method.

The main contributions of this paper lie in three aspects. First, compared with visual feature based image representations, the proposed BRCSS can alleviate the semantic gap by using semantic space based image representations. Second, compared with other semantic space based image representations, BRCSS uses the contextual semantic representations of images in an iterative way by generating semantic spaces with random sampling and re-weighting. Third, the semantic representation is also task dependent.

Although both BRCSS and BCSS [57] use the boosting strategy, they are fundamentally different for three reasons. First, BRCSS explores semantic relationships of exemplar classifiers with mixture Dirichlet distributions. The use of semantic representation helps to alleviate the semantic gap while BCSS only uses visual information for classification. Second, BRCSS re-weights the training images while BCSS treats local features differently. The aim of local feature encoding process focuses on minimizing the reconstruction error. This is different from the classification task which makes the updating of BCSS not as efficient as BRCSS. Third, the performance of BRCSS is better than BCSS not only because of the semantic relationship modeling but also because of the updating of training images.

BRCSS also differs from GraphSC [58] which explores local manifold structure with graph regularized sparse coding. GraphSC concentrates on the efficient usage of visual information while BRCSS explores the correlations of semantic representation. Both BRCSS and GraphSC are able to improve classification performances.

The rest of this paper is organized as follows. We give the related work in Section 2. The details of the proposed boosted random contextual semantic space based image representation for classification method are given in Section 3. To evaluate the effectiveness of the proposed method, we conduct image classification experiments in Section 4. Finally we conclude in Section 5.

2. Related work

Visual information had been widely used for the classification task. Sivic and Zisserman [35] used SIFT features [25] for video retrieval with good performance. There were also many other local features [1,2,6,10] that were proposed. Dalal and Triggs [6] proposed the histograms of oriented gradients (HoG) and applied it for human detection with less computational complexity compared with SIFT feature. The speeded up robust features algorithm (SURF) was proposed by Bay et. al. [2] while Fan et. al. [10] proposed the MROGH feature. Alcantarilla et. al. [1] proposed the KAZE feature. Many works had been done using these features [7,40,42,50,54,56]. Zhang et. al. [54,56] used the SIFT feature for image classification with the sparse coding technique. Datta et. al. [7] used it for image retrieval.

Although the visual based image representation had been proven very effective, it still had one problem. The visual information had no explicit semantic meanings. To solve this problem, many works had been done [4,14,20,21,23,28,30,31,37,39,41,43,45,49,51,52,55,58]. Training images were often used for semantic space construction [4,14,20,23,28,30,43,45,52,55]. Rasiwasia and Vasconcelos [30] used low-dimensional semantic spaces for scene classification while Hauptmann et. al. [14] used high-level concepts for video retrieval. Zhang et. al. [55] used exemplar classifiers for weak semantic space construction and then extended with sub-semantic space [52]. Pereira et. al. [28] used semantic queries for retrieval. Torresani et. al. [43] proposed the classesmes as a way to semantically represent images and applied it for object category recognition. Vogel and Schiele [45] used this technique for content-based image retrieval. Bosch et. al. [4] used a hybrid generative/discriminative approach for scene classification while Li and Perona [20] used a Bayesian hierarchical model. A co-clustering based method was proposed by Liu and Shah [23]. For specific tasks, using the training images for semantic space representation was very efficient. However, this strategy failed when the semantics were hard to classify. Besides, visual polysemy also hindered the discriminative power of the learned semantic representation.

In order to make use of other information along with the training data, researchers also leveraged extra data [21,31,39,49,51]. Li et. al. [21] proposed to construct ObjectBank by collecting images from the Internet while Russell et. al. [31] used LabelMe to make use of human labor for image annotation. Zhang et. al. [51] tried to learn discriminative codebooks using the information of other datasets while Yang et. al. [49] used web images for semantic video indexing. A human
labeling technique was proposed by Tang et. al. [39] for image tagging. However, once the extra data was collected, this approach would be still dataset dependent. Besides, these methods only used the visual information for initial semantic space construction. Since Internet images were often contaminated with noise, the resulting semantic spaces may be biased and cannot represent images well.

The use of attributes had also been widely studied [15–17, 36, 38]. Attributes were chosen by humans as the concepts that could be easily distinguished by computers. Kovashka and Grauman [15] tried to discover attribute shades of meanings. Kovashka et. al. [16] tried to interactively search images with attribute feedback. Shih et. al. [36] tried to learn part detectors for object recognition while Lampert et. al. [17] detected objects by transferring between classes. One problem with the attribute based image representation was that not all of the images can be represented well by simple attributes or their combinations. Besides, the visual information and attribute representation were still treated separately.

Boosting was proposed by Schapire [34] to combine a number of weak learners into a strong learner provided the weak learners are not too ‘weak’ [53]. It was widely used and extended by researchers. For example, AdaBoost [46], RankBoost [26] and random forest [5]. It was also widely used for visual classification tasks [24, 27, 54]. Liu et. al. [24] used it by combining with the LDA-based features while Paisitkriangkrai et. al. [27] proposed the Randomboost for multiclass classification. By iteratively adding weak learners, we could get reliable image classification performances. Besides, the use of random selection strategy was also effective [5, 26, 27]. This strategy can help to improve the performance with the combination scheme. The selective search technique [44] and semantic preserving model [47] were also explored by researchers.

3. Visual recognition With boosted random contextual semantic spaces

In this section, we give the details of the proposed boosted random semantic space based representation for image classification. We first extract local features from images and encode them with the sparse coding technique [56]. Max pooling is used to get visual representations of images. Classifiers are then trained with random selection to generate semantic space. We predict the classes of images using the semantic space based representations and re-weight images in a boosted way. Fig. 1 gives the flowchart of the proposed method.

3.1. Image representation with visual features

Let \( X = [x_1, \ldots, x_N] \) be the \( N \) local features with \( x_n \in \mathbb{R}^{D \times 1}, n = 1, \ldots, N \), where \( D \) is the dimension of local features. Local features are encoded by optimizing over the following problem as:

\[
[\alpha, B] = \arg\min_{\alpha, B} \sum_{n=1}^{N} || x_n^T - \alpha_n^T \times B ||^2 + \lambda \parallel \alpha_n \parallel_1
\]  

where \( B \in \mathbb{R}^{K \times D} \) is the codebook to be learned, \( K \) is the codebook size, \( \lambda \) controls the sparsity of encoding parameters. Let \( \alpha = [\alpha_1, \ldots, \alpha_N] \), this problem is often solved by alternatively keeping \( \alpha/B \) fixed. When \( \alpha \) is fixed, the codebook \( B \) can be learned by:

\[
B = \arg\min_{B} \sum_{n=1}^{N} || x_n^T - \alpha_n^T \times B ||^2
\]  

When \( B \) is fixed, each encoding parameter \( \alpha_n \) can be optimized independently as:

\[
\alpha_n = \arg\min_{\alpha_n} || x_n^T - \alpha_n^T \times B ||^2 + \lambda \parallel \alpha_n \parallel_1
\]  

which can be solved using the feature-sign-search algorithm [20].
After local features are encoded, max pooling is used to extract image representation. It extracts the maximum absolute value of each dimension of parameters within an image region. Formally, let $A$ be the number of local features within one region, the $i$-th max pooled visual representation $h_i$ of this region can then be calculated as:

$$h_i = \max(|\alpha_{1,i}|, |\alpha_{2,i}|, \ldots, |\alpha_{A,i}|)$$

(4)

The region’s representation can then be obtained as $h = [h_1; h_2; \ldots; h_K]$. Spatial pyramid matching [18] is also used to combine local feature’s spatial information. This is achieved by first iteratively dividing images into finer sub-regions ($2^L, L = 0, 1, 2$). Each sub-region is represented by max pooling the encoded local features within this sub-region. The sub-regions’ representations are then concatenated for final image representation.

### 3.2. Contextual random semantic space construction

After visual representations of images are obtained, we can use them for semantic space construction. Suppose we have $M$ training images as $(h^1, y^1), \ldots, (h^M, y^M)$, where $y^m \in \mathcal{Y} = \{1, 2, \ldots, C\}$, $m = 1, \ldots, M$ are the image labels with $C$ is the number of classes. We use superscript to indicate the index of images. We randomly select $W$ images ($W < M$) for semantic space construction and repeat this process for $S$ times. For each random selection, we train $C$ linear classifiers to separate images of different classes apart as:

$$\hat{y}^w_{c,s} = \max_{\beta_{c,s}} \beta_{c,s}^T h^w_s, \ s = 1, \ldots, S$$

(5)

where $\beta_{c,s}$ are the classifier parameters, $h_s$ is the randomly selected image’s representation, $s$ is the random selection index.

The classifiers can be trained by over-fitting by minimizing the sum of regularized empirical loss over $W$.

$$\min_{\beta_{c,s}} \sum_{w=1}^{W} \ell (\beta_{c,s}^T h^w_s, \hat{y}^w_{c,s}) + \lambda_1 \| \beta_{c,s} \|^2$$

(6)

where $\| \cdot \|^2$ is the regularization term to avoid over-fitting. $\lambda_1$ controls the relative influence of summed loss and regularization. $\hat{y}^w_{c,s} = 1$ if $\hat{y}^w_{c,s} = c$, otherwise $\hat{y}^w_{c,s}$ is set to $-1$. We use the quadratic hinge loss [19] as:

$$\ell (\beta_{c,s}^T h^w_s, \hat{y}^w_{c,s}) = [\max(\beta_{c,s}^T h^w_s \times \hat{y}^w_{c,s} - 1, 0)]^2$$

(7)

The semantic space can then be constructed using the learned classifiers. This is achieved by concatenating the predicted values of these learned classifiers. Since each classifier is trained to predict the classes of images, it bears some semantic possibility that one image belongs to a particular class. Each image can then be mapped into the semantic space as $k^c_m = [\hat{y}^1_{c,s}; \ldots; \hat{y}^S_{c,s}], m = 1, \ldots, M, s = 1, \ldots, S$. In order to combine the representation power of each randomly generated semantic space, we concatenate them together as $k^m = [k^1_m; \ldots; k^S_m]$. Besides, the semantic concepts are inherently correlated. For example, ‘sea’ and ‘sky’ often appear on the same image while ‘sunrise’ and ‘sunset’ rarely co-occurs. There are often multiple objects on the same image which may have positive responses to a number of classifiers. Moreover, the visual polysemy should also be considered. To solve these problems, we model the contextual information of semantics with mixture Dirichlet distributions as:

$$P_K|Y(k|y, A^k) = \sum_{c} \gamma_c^Y \times Dir(k; \xi_c^y)$$

(8)

where $Y$ is the random variable chosen from the semantic classes $\mathcal{Y}$, $c = 1, \ldots, C$. $\gamma_c$ is the parameter for combination and $Dir(k; \xi)$ is a Dirichlet distribution with parameters $\xi = \{\xi_1, \ldots, \xi_C\}$. $Dir(k; \xi)$ has the form as:

$$Dir(k; \xi) = \frac{\Gamma(\sum_{c=1}^{C} \xi_c)}{\prod_{c=1}^{C} \Gamma(\xi_c)} (\prod_{c=1}^{C} (k_c)^{\xi_c - 1})$$

(9)

where $\Gamma(\cdot)$ is the Gamma function. The posterior probabilities can then be calculated as:

$$P_{Y|K}(y|k) = P_{Y|K}(k|y)P(y)/P_K(k)$$

(10)

Usually, uniform class prior $P(y)$ is used. We use the posterior probabilities for the contextual semantic space based image representation as $I = (P_{Y|K}(1|k), \ldots, P_{Y|K}(C|k))^T$. The maximum likelihood estimation strategy [30] is used to learn the parameters.

### 3.3. Boosted random contextual semantic spaces for classification

Instead of directly using the learned classifiers for classification, we try to combine a number of random contextual semantic space based classification schemes for joint classification. This is because the discriminative power of one semantic space is limited. Besides, the learned visual classifiers may not be able to distinguish images well in single round. We use the boosting strategy for semantic space combination.
Algorithm 1 Training phrase of the proposed boosted random contextual semantic space based image representation for classification method.

**Input:**
The training images and labels with initial weights, boosting iteration number $J$;

**Output:**
The learned classifiers;

1: for $\text{iter} = 1, 2, \ldots, J$ 
2: Randomly select $W$ images for $S$ times. For each selection, train linear classifiers with weighted training images for semantic space construction. Map images into the semantic spaces accordingly and concentrate them for image representation as described in Section 3.2.
3: Generate the mixture Dirichlet distributions to combine the contextual correlations of the semantic concepts using the training images by maximum likelihood estimation;
4: Calculate the posterior probabilities for the contextual semantic space based image representation with uniform class prior using Eq. 10. Train the classifiers for predictions.
5: Update the predicted values and the weighting parameters using Eq. 11 and 12.
6: end for.
7: return The learned classifiers;

For the $j$-th iteration, let $\tilde{y}^m_j$ be the predicted value of the linear SVM classifier for the $m$-th image, the final predicted value of the $m$-th image is calculated as:

$$f_j(l^m) = f_{j-1}(l^m) + \tilde{y}^m_j, \quad j \geq 1, \quad f_0 = 0. \quad (11)$$

The weight can then be calculated as:

$$v_m = \exp^{-f_j(l^m) \times s m}, \quad m = 1, \ldots, M \quad (12)$$

If one image is correctly/wrongly classified, its weight will decrease/increase accordingly. In this way, we can gradually concentrate on images that are hard to classify. These weighted images are then randomly selected to generate new semantic spaces as:

$$y^w_j = \max_{c,s} \beta^T_{c,s,j} h^w_i$$

by minimizing the summed loss over training images with $L_2$ constraint as:

$$\min_{\beta_{c,s}} \sum_{w=1}^W v_w \ell(c, s, j, y^{w,j} c, s, j) + \lambda_1 \| \beta_{c,s,j} \|^2 \quad (14)$$

The mixture Dirichlet distributions for the $j$-th iteration can then be modeled as:

$$p(k^{\ell} | y, \Lambda^{k, j}) = \sum_c y^{c \ell}_c \times \text{Dir}(k^{\ell} ; \xi^{c \ell}_c) \quad (15)$$

with:

$$\text{Dir}(k^{\ell} ; \xi^{c \ell}_c) = \Gamma\left(\sum_{c=1}^C \xi^{c \ell}_c + 1\right) \prod_{c=1}^C \Gamma\left(y^{c \ell}_c + 1\right) \prod_{c=1}^C \Gamma\left(\xi^{c \ell}_c + 1\right) \quad (16)$$

In this way, we can get the new contextual semantic space based representations for classifier training. The predicted values are then used to re-weight training images for the next iteration. We repeat this process for $J$ times. Algorithm 1 gives the details of the training phase of the proposed boosted random contextual semantic space based image classification method.

4. Experiments

To evaluate the effectiveness of the proposed method, we conduct image classification experiments on three public image datasets: the Natural Scene dataset [18], the COREL-5000 dataset [8] and the Caltech-256 dataset [13].

4.1. Experimental setup

For each image, we densely extract local features of multi-scales with overlap, as [30,48,56]. The smallest local region is set to $16 \times 16$ pixels with an overlap of 6 pixels. Sparse coding [48] is used for codebook generation and local feature encoding. We normalize the extracted local features with $L_2$ norm. About 100,000 local features are randomly selected for codebook generation. After the codebook is learned, we fix it for local feature encoding. Spatial pyramid matching with three pyramids ($L = 0, 1, 2$) is used to combine the spatial information of local features. $\lambda$ is the parameter for sparsity control. We
can be fixed to be $0.3 \sim 0.4$. We follow the experimental setup of other researchers and use the reported results directly instead of re-implementing them for fair comparison. Images are randomly selected for classifier training and the other images are used for performance evaluation. This process is repeated for several times to get reliable results. The performance is evaluated using the mean classification rates.

### 4.2. Natural scene dataset

This dataset has fifteen natural scenes. It is initially proposed by Li and Perona [20] and Bosch et. al [4]. There are about 200 to 400 images per class with the average size of $300 \times 250$ pixels. We randomly select 100 images per class for training and use the rest of images for evaluation. This process is repeated for ten times.

We give the performance comparisons with other methods in Table 1. We can have several conclusions from Table 1. First, compared with visual feature based strategies [12,18,48], the use of semantic based representation can help to alleviate the semantic discrepancy. Hence BRCSS can improve over them by at least 5%. Second, compared with other semantic methods [4,18,21,29,30,52,55], BRCSS can make use of the discriminative power of learned classifiers. Besides, by combining a series of contextual semantic spaces, we can improve the classification performance over single semantic space. Specially, BRCSS outperforms Contextual Models [29] by about 10% and S3R (sub-semantic space representation) [52] by 4%. Third, BRCSS also outperforms ObjectBank [21] which uses Internet images for semantic representation. This is because Internet images are contaminated with noise which makes the learned classifiers unreliable. Finally, BRCSS improves the performance over BCSS by about 2.4%. The use of boosting strategy with the specific classification task also makes the representations task dependent, hence can improve the final classification performance.

### 4.3. COREL-5000 dataset

The COREL-5000 dataset consists of images extracted from 50 Corel CDs with each CD having 100 images. There are 1–5 labels for each image in this dataset. We follow the experimental setup as [29] and choose the first 90 images for training and use the other images for testing. Table 2 shows the performance comparisons of the proposed method with other methods.

Similar conclusions can be drawn from Table 2 as from Table 1. Compared with visually based methods [18,29], BRCSS can explore the semantic correlations which are more consistent with human perception. Besides, pLSA [4] and LDA [30] are

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Classification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScSPM [48]</td>
<td>80.28 ± 0.93</td>
</tr>
<tr>
<td>SPM [18]</td>
<td>81.40 ± 0.50</td>
</tr>
<tr>
<td>Kernel codebook [12]</td>
<td>76.67 ± 0.39</td>
</tr>
<tr>
<td>Semantic space [30]</td>
<td>73.95 ± 0.74</td>
</tr>
<tr>
<td>Contextual models [29]</td>
<td>77.20 ± 0.39</td>
</tr>
<tr>
<td>ObjectBank [21]</td>
<td>80.9</td>
</tr>
<tr>
<td>WSR-EC [55]</td>
<td>81.54 ± 0.59</td>
</tr>
<tr>
<td>S3R [52]</td>
<td>83.72 ± 0.78</td>
</tr>
<tr>
<td>BCSS [57]</td>
<td>85.42 ± 0.64</td>
</tr>
<tr>
<td><strong>BRCSS</strong></td>
<td><strong>87.85 ± 0.54</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithms qua</th>
<th>Classification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>pLSA [4]</td>
<td>40.2</td>
</tr>
<tr>
<td>LDA [30]</td>
<td>31.0</td>
</tr>
<tr>
<td>SPM [18]</td>
<td>48.4</td>
</tr>
<tr>
<td>Appearance models [29]</td>
<td>53.6</td>
</tr>
<tr>
<td>Contextual models [29]</td>
<td>57.8</td>
</tr>
<tr>
<td>BCSS [57]</td>
<td>61.7</td>
</tr>
<tr>
<td><strong>BRCSS</strong></td>
<td><strong>63.5</strong></td>
</tr>
</tbody>
</table>
Table 3
Performance comparison on the Caltech-256 dataset.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>15 training</th>
<th>30 training</th>
<th>45 training</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSMP [48]</td>
<td>23.34 ± 0.42</td>
<td>29.51 ± 0.52</td>
<td>-</td>
</tr>
<tr>
<td>ScKSMP [48]</td>
<td>27.73 ± 0.51</td>
<td>34.02 ± 0.35</td>
<td>37.46 ± 0.55</td>
</tr>
<tr>
<td>Classerem [43]</td>
<td>-</td>
<td>36.00</td>
<td>-</td>
</tr>
<tr>
<td>ObjectBank [21]</td>
<td>-</td>
<td>39.00</td>
<td>-</td>
</tr>
<tr>
<td>WSRC-EC [55]</td>
<td>35.28 ± 0.65</td>
<td>42.01 ± 0.47</td>
<td>45.82 ± 0.54</td>
</tr>
<tr>
<td>SfR [52]</td>
<td>37.85 ± 0.48</td>
<td>43.52 ± 0.44</td>
<td>46.86 ± 0.63</td>
</tr>
<tr>
<td>NBNN [3]</td>
<td>30.45</td>
<td>38.18</td>
<td>-</td>
</tr>
<tr>
<td>KSMP [13]</td>
<td>-</td>
<td>34.10</td>
<td>-</td>
</tr>
<tr>
<td>Kernel Codebook [12]</td>
<td>-</td>
<td>27.17 ± 0.46</td>
<td>-</td>
</tr>
<tr>
<td>LsSMP [11]</td>
<td>30.00 ± 0.14</td>
<td>35.74 ± 0.10</td>
<td>38.54 ± 0.36</td>
</tr>
<tr>
<td>BCSS [57]</td>
<td>39.75 ± 0.58</td>
<td>45.10 ± 0.49</td>
<td>48.04 ± 0.51</td>
</tr>
<tr>
<td>BRCSS</td>
<td><strong>42.58 ± 0.60</strong></td>
<td><strong>47.25 ± 0.51</strong></td>
<td><strong>49.92 ± 0.47</strong></td>
</tr>
</tbody>
</table>

designed for text analysis. However, the semantic discrepancy of visual features is larger and unpredictable compared with text. Moreover, the combination of semantic spaces makes it more reliable and discriminative. The joint modeling of classification task with image representation is also useful for image class prediction. Hence BRCSS can outperform contextual/appearance models by 5.7/9.9% respectively. The combination of a series of semantic spaces also makes the proposed method robust to noise as we can gradually concentrate on images that are hard to classify. Furthermore, BRCSS is able to outperform BCSS by about 1.8%. This again shows the effectiveness of the proposed method.

4.4. Caltech-256 dataset

The Caltech-256 dataset has 29,780 images of 256 classes with large variability. There are at least 80 images for each class. We randomly select 15/30/45 training images per class for performance evaluation and use the rest images for testing. This process is repeated for ten times. Table 3 gives the performance comparisons of BRCSS with other methods.

We can see from Table 3 that the semantic-based method can achieve comparable or superior performances compared with visually based methods. This shows the usefulness of using semantic representations. Besides, the use of discriminative classifiers often outperforms generative classifiers. This is because discriminative classifiers can learn more separable boundary. Besides, generative classifiers often assume pre-defined distributions which cannot be justified by real data. The proposed BRCSS method dynamically combines discriminative and generative classifiers, hence can get more discriminative image representations. Moreover, the encoding of local features also makes it more discriminative than directly using local features [3]. The results on the Caltech-256 dataset again demonstrate the effectiveness of the proposed BRCSS method.

Compared with the performances on the Natural Scene dataset and the COREL-5000 dataset, the relative improvement of BRCSS over visual only based methods [48] increases. We believe this is for two reasons. First, images of the Caltech-256 dataset often have various objects while images of the other two datasets not. This makes the contextual semantic space based image representation be able to model the mutual relationships of different semantics better. Second, with the increment of classes, the discriminative power of visual based classifiers decreases. However, by using semantic based representation, we can alleviate this drawback. Besides, the relative improvements over semantic space based methods [29] also increase. This is because the boosting strategy can take advantages of a series of semantic spaces. Moreover, BRCSS also improve the performance over BCSS on the Caltech-256 dataset.

The proposed method can be combined with various visually based representation methods (e.g. the Fisher Vector [32] and the convolutional neural network [9]). By combining the Fisher Vector, we can achieve 51.6% accuracy when 30 training images are used on the Caltech-256 dataset. This exceeds [32] by 4.2%. Similarly, when combined with the convolutional neural network [9], we can improve the performance to 53.2%.

4.5. Parameter influences

The codebook size is an important parameter which influences the final classification performance. A smaller codebook may not be able to separate local features well while a larger codebook costs more computational power and memory. To show the influences of codebook size, we give the classification accuracy changes with the codebook size in Fig. 2 on the three datasets.

We can see from Fig. 2 that the performance increases with the codebook size. However, the relative improvement decreases. This means choosing a proper codebook size is important for final classification. We use 1024 as the codebook size in this paper.

The number of Dirichlet distributions influences the discriminative power of each iteration. If it is too small, the resulting image representation may have no discriminative power at all. However, if this number is too large, we may represent images too ‘finely’. To show the influences of the number of Dirichlet distributions, we give the performance changes in Fig. 3. We can see from Fig. 3 that choosing a proper number is important for reliable performances. Besides, we can see that proper
The number of boosting iterations is also important for the performances. For each iteration, we can reduce the training errors by classifying images with the random contextual semantic spaces. Besides, the hard to classify images are re-weighted for the next iteration. In this way, we can gradually improve the classification accuracy. We plot the performance changes with the number of boosting iterations in Fig. 4. We can see the classification performance increases with the number of iterations. By adding more and more contextual semantic space based representations guided by the classification task, we can gradually improve the performances. Besides, since the three datasets have different difficulties, the optimal numbers of iterations also vary. However, the final performances become statable as long as the number of iterations is enough.
The number of training images $W$ and the number of random selection times $S$ are other two parameters which influence the performances. Figs. 5 and 6 give the influences of the number of training images $W$ and the number of random selection times $S$ respectively. We can see from Fig. 5 that the performance increases with the number of training images. Besides, the influence of training images' number is relatively less important as we can iterate more times for compensation. As to the influences of random selection times, we can see from Fig. 6 that the performance increases with the selection times at first. However, with enough selection times, the performance becomes stable. Besides, the Caltech-256 dataset requires more selection times as it is relatively difficult to classify than the other two datasets.
5. Conclusion

In this paper, we proposed a novel image classification method by using the boosted random contextual semantic space based image representation. The BRCSS dynamically combined the visual and semantic representations into a unified process. The visual representations were used for initial semantic space representations and then re-weighted to generate the corresponding semantic spaces iteratively. During each iteration, classifiers were adjusted to focus on images that were hard to classify. In this way, we can generate a series of semantic spaces for representations which suited the classification task. Experimental results on three image datasets proved the effectiveness of the proposed method.

Acknowledgement

This work is supported by National Natural Science Foundation of China: 61303154. This work is partially supported by National Natural Science Foundation of China: 61332016, National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences.

References
