Use Bin-Ratio Information for Category and Scene Classification

Nianhua Xie[1,2], Haibin Ling[2], Weiming Hu[1], Xiaoqin Zhang[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China

[2]Dept. of Computer and Information Science, Temple University, Philadelphia, USA

{nhxie, wmhu, xqzhang}@nlpr.ia.ac.cn, {hbling}@temple.edu

Abstract

In this paper we propose using bin-ratio information, which is collected from the ratios between bin values of histograms, for scene and category classification. To use such information, a new histogram dissimilarity, bin-ratio dissimilarity (BRD)*, is designed. We show that BRD provides several attractive advantages for category and scene classification tasks: First, BRD is robust to cluttering, partial occlusion and histogram normalization; Second, BRD captures rich co-occurrence information while enjoying a linear computational complexity; Third, BRD can be easily combined with other dissimilarity measures, such as L_1 and χ^2, to gather complimentary information. We apply the proposed methods to category and scene classification tasks in the bag-of-words framework. The experiments are conducted on several widely tested datasets including PASCAL 2005, PASCAL 2008, Oxford flowers, and Scene-15 dataset. In all experiments, the proposed methods demonstrate excellent performance in comparison with previously reported solutions.*

1. Introduction

Histogram-based representation is very popular in many computer vision tasks since it captures rich information that leads to powerful discriminability. In particular, it is widely used in category and scene classification, such as in the bag-of-words models [8, 9, 21, 32, 37, 34]. In these models, an image is represented by the histogram (frequency) of vector quantized visual words. These histograms are then combined with classifiers (e.g., support vector machines) for classification. Many studies have been focusing on different aspects of the framework, including the generation of visual dictionary, quantization of visual patches into visual words, efficient and effective inference on the histogram representations, etc. In this paper, we focus on the study of dissimilarity measures between such histogram representations.

There are two major motivations for our study: *partial matching* and *co-occurrence information*. Partial matching
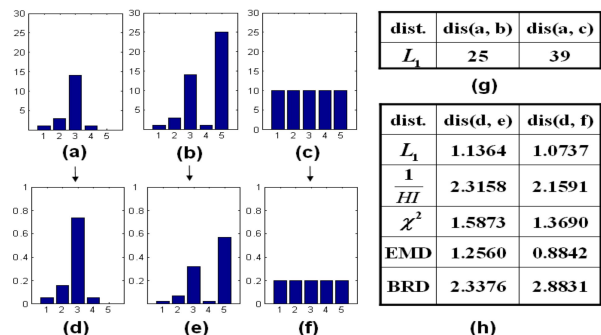

Figure 1. An example of the partial matching and histogram normalization problem. (a), (b) and (c) show three histograms before normalization. Histograms in (a), (b) and (c) are [1 3 14 1 0], [1 3 14 1 25] and [10 10 10 10 10] respectively. (d), (e) and (f) show the corresponding normalized histograms. (g) shows the L_1 distances between (a) and (b), (a) and (c) before normalization. (h) shows dissimilarities between (d) and (e), (d) and (f) calculated by several different methods, where "HI", "EMD", and "BRD" indicates "Histogram Intersection", "Earth Mover's Distance", and the proposed "bin-ratio dissimilarity" respectively.

is an important issue in category and scene classification, because images often contain a large amount of clutters that are irrelevant to the object-of-interest. Such clutter information unavoidably brings noises into histogram representations and causes troubles for inferences. A toy example is shown in Figure 1 where (a) and (b) represent histograms of images from the same category. While bins 1 to 4 in (a) and (b) are exactly the same, bin (5) is very different that simulates large background clutters in the image where (b) is computed. Although it is known that some histogram comparison methods (e.g. L_1, Histogram Intersection [31], Earth Mover's distance (EMD) [27]) are robust to partial matching, most of them meet problems when histograms are normalized. For example, in Figure 1 (h), these methods treat (d) and (f) as more similar than (d) and (e).

In histogram-based representation, co-occurrence captures the correlation information between pairs of histogram bins (i.e., joint frequency of visual words in the bag-of-words model). Many existing histogram measures implicitly assume independent distribution of different visual words, however, such assumption is often violated in the real world. For example, a word describes "eye" and a

word describes "mouth" usually appear together for face images. This observation has triggered research efforts [30, 2, 28, 19, 35, 1, 17, 14, etc.] that take into account joint word distributions for improving final inference. Different than previous work, in this paper, we use the co-occurrence information which directly comes from histogram (i.e. not from spatial space), for improving dissimilarity measures between histograms.

To address the above two issues, we propose using *bin-ratio information* for category and scene classification. Bin-ratios are defined as the ratios between bin values of histograms. Specifically, given a n-bin histogram $\mathbf{h} \in \mathbb{R}^n$ (e.g., a word frequency used in the bag-of-words model), we define its *ratio matrix* as $H = (\mathbf{h}(j)/\mathbf{h}(i)) \in \mathbb{R}^{n \times n}$. The ratios defined this way are robust to normalization and cluttering as well as capture co-occurrence information between all pairs of bins. Then, given two histograms, we define their *bin-ratio dissimilarity* (BRD) as the sum of squares of normalized differences over all matrix elements. As shown in Figure 1(h), the new dissimilarity handles simultaneously partial matching and normalization problems. Furthermore, we show that the bin-ratio dissimilarity has a linear instead of quadratic computational complexity. Finally, the bin-ratio dissimilarity is combined with other distance measures (e.g., L_1 and χ^2) for category and scene classification tasks. Our experiments are conducted on several widely tested datasets including PASCAL 2005, PASCAL 2008, Oxford flowers, and Scene-15. In all experiments, our methods demonstrate excellent performance in comparison with previously reported solutions.

In summary, our main contribution lies in using the bin-ratio information for category and scene classification. Specifically, we introduce bin-ratio dissimilarity that has the following advantages: First, it is robust to background clutter and histogram normalization; Second, it captures co-occurrence information while keeps an linear computational complexity; Third, the method is flexible and can be easily combined with different distances that usually provide complimentary information. Furthermore, the proposed methods have the potential to be used for a wide range of tasks, since bin-ratio dissimilarity is general to all histogram representations.

The rest of the paper is organized as follows. §2 reviews the related work. §3 presents the proposed BRD measure and how it is combined with other methods for category and scene classification. §4 experimentally compares ratio distance against other methods on several widely tested datasets. Finally, §5 concludes the paper.

2. Related Work

Category and scene classification is a very active area in recent years and has attracted large amount of research efforts [25]. Our work is closely related to the histogram-based representations, especially to those under the bag-of-words framework [8, 9, 21, 32, 37, 34, etc.]. In particular, our work is directly compared with previous work where different histogram dissimilarity schemes are used [11, 21, 32, 37, 19, 22, 23, 33].

Due to the increasing popularity of histogram based descriptors, reliable and efficient histogram comparisons become an important research topic. Aside from the classic "bin-to-bin" distances (e.g., L_1, L_2, χ^2), "cross-bin" distances [27, 18, 12, etc.] have been recently drawing attentions by many researchers. These distances allow cross bin comparison between two histograms to gain robustness to histogram distortions. In the following, we review some popular bin-to-bin and cross-bin distances that are widely used in visual category recognition, as these distances are most related to our study.

The most simple form of bin-to-bin distance is the Minkowski-form distance, including L_1 and L_2 distances. Recently, the Histogram Intersection introduced by Swain and Ballard [31] has shown good performance and efficiency in visual category recognition [15, 20, 36]. The χ^2 distance [26] has been widely used in visual category recognition [8, 21, 22, 32, 37], because of its simplicity and good performance. Although these widely used bin-to-bin distances have achieved good performances, they can not handle partial matching with normalized histograms, for the normalization will greatly change the value difference between corresponding bins, as shown in Fig. 1. Compared with these bin-to-bin distances, the proposed bin-ratio dissimilarities are more robust for normalized partial matching and has a comparable complexity. Furthermore, bin-ratio dissimilarity contains high order information, which is very useful in classification.

Rubner et al. [27] propose a cross-bin distance called the Earth Mover's Distance (EMD). EMD defines the distance computation between distributions as a transportation problem, which means EMD can do matching cross different bins. Zhang et al. [37] show outstanding performance of EMD on various datasets. However, the time complexity of EMD is larger than $O(N^3)$, where N is the number of histogram bins. In addition, as shown is Fig. 1, EMD also suffers from the normalized partial matching, because it depends on the value differences between bins. Our work shares the philosophy by including cross-bin interactions. In our methods, however, the interactions are between bins from the same histogram, while interactions in cross-bin distances act on bins from different histograms. Moreover, the proposed dissimilarity measures are computationally more efficient than most cross-bin distances.

Co-occurrence statistics is known to be relevant for category classification. Many studies have been conducted toward modeling spatial co-occurrence of visual words [30, 2, 28, 19, 35, 1, 17, 14, etc.]. Agarwal and Triggs [1] propose

a hyperfeature which exploits spatial co-occurrence statistics of features. Li et al. [17] introduce a general framework called Markov stationary features (MSF), which characterizes the spatial co-occurrence of histogram patterns by Markov chain models. In comparison, the proposed BRD focuses on the co-occurrence in feature space. We use the information directly comes from histograms, therefore the bin-ratio dissimilarities are more efficient and readily to be used in different scenarios.

Our work is also motivated by recent matching algorithms with high-order statistics, such as the second-order programming [2] and matching [16] and high-order matching [6]. In [6], the authors propose a tensor-based algorithm for high-order graph matching. They formulate the hypergraph matching problem as the maximization of a tensor function over all permutation of the features tuples. Our work is partially inspired by this idea in that the proposed dissimilarity measures use the high-order information over different bins within a histogram.

3. Bin-Ratio Dissimilarity between Histograms

Notation: In the rest of the paper, we use a bold letter to indicate a vector, e.g., $\mathbf{h}$, and denote its ith elements as $\mathbf{h}(i)$, or h_i when there is no ambiguity. We use $\| \cdot \|_1$ and $\| \cdot \|_2$ to indicate L_1 and L_2 norms respectively.

3.1. Bin-Ratio Dissimilarity

Let us denote a normalized histogram $\mathbf{h} \in \mathbb{R}^n$ with n bins in total. In this paper, we assume L_2 normalization is applied to histograms in our proposed methods, i.e.,

$$\| \mathbf{h} \|_2^2 = \sum\nolimits_{1 \le k \le n} h^2(k) = 1 . \tag{1}$$

To capture pairwise relationship between histogram bins, we define *ratio matrix* $H \in \mathbb{R}^{n \times n}$ of a histogram $\mathbf{h}$ such that the matrix elements capture the ratios of paired bin values. In particular, $H \in \mathbb{R}^{n \times n}$ is defined as

$$H = \left(\frac{h_j}{h_i} \right)_{i,j} = \begin{pmatrix} \frac{h_1}{h_1} & \frac{h_2}{h_1} & \frac{h_3}{h_1} & \cdots & \frac{h_n}{h_1} \\ \frac{h_1}{h_2} & \frac{h_2}{h_2} & \frac{h_3}{h_2} & \cdots & \frac{h_n}{h_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{h_1}{h_n} & \frac{h_2}{h_n} & \frac{h_3}{h_n} & \cdots & \frac{h_n}{h_n} \end{pmatrix} \tag{2}$$

This ratio matrix is invariant to normalization since each element is invariant to normalization. Furthermore, each matrix element h_j/h_i measures the relation between elements h_i, h_j in the original histogram $\mathbf{h}$. This inspires us to compare ratio matrices when comparing two histograms. Intuitively, the ith row of the ratio matrix represents the ratios all bin values to the value in the ith bin. We deal with the ith row of a ratio matrix instead of the ith bin of a histogram. Specifically, to compare the ith bins of two histograms $\mathbf{p}$ and $\mathbf{q}$, we can define the dissimilarity d_i as the

sum of squared differences (SSD) between i^{th} rows of corresponding ratio matrices, that is,

$$d_i(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{n} \left(\frac{q_j}{q_i} - \frac{p_j}{p_i} \right)^2 \tag{3}$$

Such definition apparently suffers from the instability problem when p_i and q_i are close to zero. To avoid this problem, we include the normalization part and define the following dissimilarity

$$d_{br,i}(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{n} \left(\frac{\frac{q_j}{q_i} - \frac{p_j}{p_i}}{\frac{1}{q_i} + \frac{1}{p_i}} \right)^2 \tag{4}$$

This normalization makes the whole dissimilarity variant to normalization. However, the gaining in robustness is more important, and strict normalization invariant is not necessary. Consequently, we define the *bin-ratio dissimilarity* (BRD) $d_{br}(\mathbf{p}, \mathbf{q})$ between two histograms $\mathbf{p}, \mathbf{q}$ as

$$d_{br}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} d_{br,i}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\frac{q_j}{q_i} - \frac{p_j}{p_i}}{\frac{1}{q_i} + \frac{1}{p_i}} \right)^2 \tag{5}$$

Though seems to have quadratic complexity, BRD can actually be computed in linear time, as derived in the following. Starting from (4), $d_{br,i}$ can be rewritten as

$$
\begin{aligned}
d_{br,i}(\mathbf{p}, \mathbf{q}) &= \frac{\sum_{j=1}^{n}(q_j p_i - q_i p_j)^2}{(p_i + q_i)^2} \\
&= \frac{\sum_{j=1}^{n}(q_j^2 p_i^2 + q_i^2 p_j^2 - 2q_j p_i q_i p_j)}{(p_i + q_i)^2} \\
&= \frac{p_i^2 + q_i^2 - 2p_i q_i \sum_{j=1}^{n} p_j q_j}{(p_i + q_i)^2} \tag{6} \\
&= \frac{(p_i + q_i)^2 - 2p_i q_i(1 + \sum_{j=1}^{n} p_j q_j)}{(p_i + q_i)^2} \\
&= \frac{(p_i + q_i)^2 - p_i q_i \sum_{j=1}^{n}(p_j + q_j)^2}{(p_i + q_i)^2} \tag{7} \\
&= 1 - \frac{p_i q_i}{(p_i + q_i)^2} \|\mathbf{p} + \mathbf{q}\|_2^2 . \tag{8}
\end{aligned}
$$

Note that the L_2 normalization (1) assumption is used in steps (6) and (7), though the L_2 normalization is not critical for deriving the linear complexity. In other words, other normalization schemes may also yield a linear computation formula. Also note that we define $d_{br,i}(\mathbf{p}, \mathbf{q})$=0 for the degenerated case where p_i=q_i=0[1], though in the derivation we ignore such case for clarity. Using formulation (8), the BRD d_{br} can be written as

$$
\begin{aligned}
d_{br}(\mathbf{p}, \mathbf{q}) &= \sum_{i=1}^{n} \left(1 - \frac{p_i q_i}{(p_i + q_i)^2} \|\mathbf{p} + \mathbf{q}\|_2^2 \right) \\
&= n - \|\mathbf{p} + \mathbf{q}\|_2^2 \sum_{i=1}^{n} \frac{p_i q_i}{(p_i + q_i)^2} . \tag{9}
\end{aligned}
$$

---

[1]We can alternatively add a very small value in the denominator of (7). Compared with multiplication, the adding complexity can be ignored.

Using (9), BRD can be calculated in linear time. This is because both items $\|\mathbf{p} + \mathbf{q}\|_2^2$ and $\sum_{i=1}^{n} \frac{p_i q_i}{(p_i + q_i)^2}$ have linear complexities and the combination of them take only constant time.

3.2. Combine Bin-Ratio Dissimilarity with Other Dissimilarities

While robust for partial matching and normalization, BRD suffers from problems caused by small noisy bins. Specifically, from equation (8), we see that when one of p_i, q_i is zero and the other is a small noise, we have $d_{br,i}(\mathbf{p}, \mathbf{q}) = 1$. This observation implies that BRD can be sensitive to small noises. On the other hand, many classical histogram measures handle nicely small noises. Therefore, instead of directly using BRD, we combine it with other methods for complementary gaining.

We first combine BRD with the widely used L_1 distance, which is known to be robust to outliers small noises. We use bin-level combination (different than that in the multiple-kernel methods [10, 13]), since intuitively this leads to closer interaction between the two methods. The combined dissimilarity, which we call L_1-*BRD*, is derived from (8) as following

$$d_{l_1 br,i}(\mathbf{p}, \mathbf{q}) = |p_i - q_i| d_{br,i}$$

$$= |p_i - q_i| - \frac{|p_i - q_i| p_i q_i}{(p_i + q_i)^2} \|\mathbf{p} + \mathbf{q}\|_2^2 \quad (10)$$

$$d_{l_1 br}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} d_{l_1 br,i}(\mathbf{p}, \mathbf{q})$$

$$= \|\mathbf{p} - \mathbf{q}\|_1 - \|\mathbf{p} + \mathbf{q}\|_2^2 \sum_{i=1}^{n} \frac{|p_i - q_i| p_i q_i}{(p_i + q_i)^2} \quad (11)$$

Intuitively, we see from Eqn. (10) that when one of bins p_i, q_i is zero, L_1-BRD automatically adopts L_1 distance, which is more robust than the original BRD. In our experiments we found that such combination yields very promising classification results (See §4).

Similar to L_1-BRD, χ^2 distance can also be combined with BRD. This combination generates the χ^2-*BRD*, which is presented below,

$$d_{\chi^2 br,i}(\mathbf{p}, \mathbf{q}) = 2 \frac{(p_i - q_i)^2}{p_i + q_i} d_{br,i}(\mathbf{p}, \mathbf{q}) \quad (12)$$

$$d_{\chi^2 br}(\mathbf{p}, \mathbf{q}) = d_{\chi^2}(\mathbf{p}, \mathbf{q}) -$$

$$2 \|\mathbf{p} + \mathbf{q}\|_2^2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2 p_i q_i}{(p_i + q_i)^3} , \quad (13)$$

where $d_{\chi^2}(\mathbf{p}, \mathbf{q}) = 2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i + q_i}$ is the χ^2 distance between $\mathbf{p}, \mathbf{q}$.

Similar to BRD, both L_1-BRD and χ^2-BRD have linear computational complexities, which make them an efficient solutions to large scale tasks.

Figure 2. Example images from Scene-15 dataset, one per class.

3.3. Kernel-based Classification

To use bin-ratio information for classification tasks, we combine BRDs with the standard bag-of-words model. To compare with other methods, such as χ^2, EMD, etc., we follow the same kernel-based framework as in [37]. For this purpose, we first build BRD kernels using the extended Gaussian kernels [5],

$$K(\mathbf{p}, \mathbf{q}) = \exp\left(-\frac{1}{A} D(\mathbf{p}, \mathbf{q})\right) , \quad (14)$$

where $D(\mathbf{p}, \mathbf{q})$ is a dissimilarity if $\mathbf{p}$ and $\mathbf{q}$ are histograms of images; A is a scaling parameter that can be determined by cross-validation. In practice, we have tested BRD, L_1-BRD, χ^2-BRD for the dissimilarity $D(.)$, and found that L_1-BRD generates the best performance.

We have not proved whether the BRD-based kernels are Mercer kernels or not. Nevertheless, during our experiments, these kernels have always generated positive definite Gram matrices. It should be noted that some widely used kernels, e.g. EMD-kernel, also have not been proofed to be Mercer kernel [37]. In addition, even some non-Mercer kernels usually work well in real applications [5].

4. Experiments

We apply the proposed BRDs (i.e., BRD, L_1-BRD and χ^2-BRD) to category and scene classification tasks on four public datasets: Scene-15 dataset [15], PASCAL VOC 2008 [7], PASCAL VOC 2005 Test2 [8], and Oxford Flowers [22]. We first compare BRDs against χ^2, L_1, and histogram intersection on Scene-15 dataset. Then we will show our results on PASCAL 2008 and 2005, by comparing directly with with the winners' results. Finally, we test L_1-BRD on the Oxford flowers and report better classification rate than all previous tested methods.

4.1. Scene-15 dataset

The Scene-15 dataset [15] (see Fig. 2 for examples) is compiled from several datasets from earlier studies [9, 24, 15]. It contains 15 categories and 4485 scene images, with 200 to 400 images per class.

For comparison, we follow the experimental setup of van Gemert et al. [11]. The classification is repeated for 10 times, with randomly split training testing sets. In [11], Histogram Intersection (HI) is used on this dataset and a
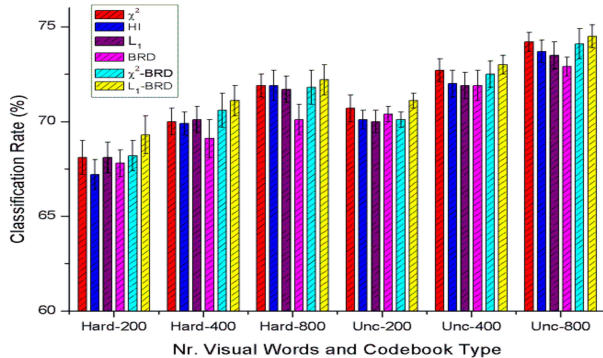
Figure 3. Classification performance results of various types of histogram similarity and dissimilarities for the Scene-15 dataset over different vocabulary sizes and codebook type. The "Hard" and "Unc" represent using the codebook type of hard assignment and codeword uncertainty respectively.

Figure 4. Example images from PASCAL 2008, one per class.

kernel codebook technique is used in comparison with standard codebook (i.e., hard codebook). We re-implement HI kernel and apply χ^2, L_1 and the proposed BRDs on this dataset with the same experimental setup for fair comparison. Specifically, in each round, a codebook vocabulary is created using k-means on the train set. Then we use normal codebook (hard assign) and kernel codebook (codeword uncertainty) respectively. For each type of codebook, we use an SVM with several histogram distances/kernels: χ^2, HI, L_1, and BRDs. The BRDs are based on the L_2 normalization, while the other distances are all based on the L_1 normalization. For multi-class classification, we use Libsvm [4] and one-versus-all scheme. Five-fold cross-validation is used on the train set to tune parameters. We report the average accuracies over 10 rounds. For each round, the accuracy is calculated by averaging the accuracies of each category.

For image features, we use SIFT descriptors sampled on a regular grid with spacing of eight pixels. SIFT descriptors are calculated on 16×16 patches. Our re-implementation results of HI are similar to the results of [11][2].

Fig. 3 shows the results of χ^2, HI, L_1, BRD, χ^2-BRD, and L_1-BRD with various codebook sizes: 200, 400, 800 and with normal codebook (hard assign) or kernel codebook (codeword uncertainty). From the results, we can see that BRD by itself performs comparable to previous tested methods already. This suggests that bin-ratios contain rich discriminative information. Furthermore, when combined with L_1 distance, the L_1-BRD generates the best performance, which validates our conjecture in §3.2.

4.2. PASCAL VOC 2008

The PASCAL Visual Object Classes Challenge [8, 7] (see Fig. 4 for examples) provides a yearly benchmark for
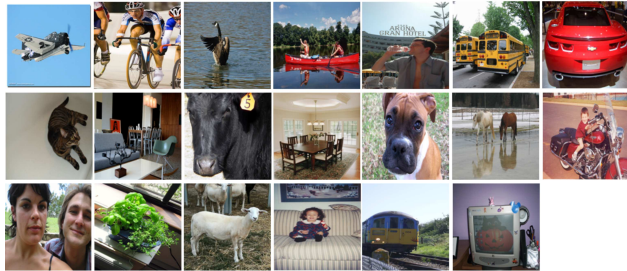
---

[2]In [11], the experimental results are summarized as visually in barcharts instead of detailed rates. Our re-implementation generates classification rates that are visually similar to theirs.

evaluating category classification methods. The PASCAL VOC 2008 consists of 20 classes with 8465 images. The whole dataset is divided into a predefined "trainval" set with 4332 images and "test" set of 4133 images. The "trainval" dataset is further divided into two parts: a validation set containing 2221 images and a training set containing the rest. Category labels are only released for "trainval" set images to avoid parameter overfitting. The results on the test set are provided by the PASCAL organizers. Therefore, the performance on the test set is more reliable.

We follow the framework of the winners of PASCAL 2007 by Marszalek et al. [21] and PASCAL 2008 by Tahir et al. [32]. Compared with the winner of PASCAL 2008, the main difference lies in that L_1-BRD is used instead of χ^2. Specifically, for each image we use two sampling methods for local patches: Harris-Laplace point sampling and densely sampling with every six pixels. Then each image patch is further represented by SIFT and four color-SIFT descriptors [29]: OpponentSIFT, rgSIFT, C-SIFT and RGB-SIFT. These color-SIFT descriptors have specific invariance properties and can be used to improve classification performance [29]. The patch descriptors of training images are clustered by k-means to generate a vocabulary of 4000 words. Kernel codebook (codeword uncertainty) is used to for further improvement. After that, three different spatial pyramids are used: the whole image without subdivision (1x1), image parts divided into 4 quarters (2x2) and a pyramid with three horizontal bars (1x3). L_1-BRD is used to calculate histogram dissimilarities followed by the extended Gaussian kernel (14) to be used in SRKDA [3] as in the winner's method. Parameters are calculated on validation set and further used for test set.

Tables 1 lists the results of our method against the winner's method [32] on both validation and test sets. The results show that, on the validation set where groundtruth is known, L_1-BRD outperforms the winner's method in 14 out of 20 categories, as well as the average performance. On the test set, L_1-BRD works not only better than the winner's method, but also outperforms the best achieved results (from different methods) in 11 out of 20 categories. Since we strictly follow the winner's method except histogram

	Validation Set		Test Set		
	Winner [32]	L_1-BRD	Winner [32]	Best achieved [7]	L_1-BRD
Aeroplane	**79.2**	77.5	79.5	**81.1**	79.7
Bicycle	39.7	**44.6**	54.3	54.3	**56.3**
Bird	49.1	**50.8**	61.4	**61.6**	61.1
Boat	62.6	**63.0**	64.8	**67.8**	66.5
Bottle	18.8	**20.1**	30.0	30.0	**30.6**
Bus	52.3	**55.1**	52.1	52.1	**56.5**
Car	**55.5**	54.8	**59.5**	**59.5**	58.9
Cat	**56.2**	55.1	59.4	**59.9**	58.1
Chair	42.5	**45.4**	48.9	48.9	**49.4**
Cow	24.1	**26.4**	33.6	33.6	**34.9**
Dining table	25.7	**29.0**	37.8	40.8	**43.5**
Dog	32.8	**36.8**	46.0	**47.9**	47.0
Horse	47.1	**49.5**	66.1	67.3	**67.5**
Motorbike	39.9	**41.5**	64.0	**65.2**	62.9
Person	**88.0**	87.2	86.8	**87.1**	86.6
Potted plant	24.1	**28.5**	29.2	31.8	**33.2**
Sheep	30.8	**31.2**	42.3	42.3	**42.7**
Sofa	38.7	38.7	44.0	45.4	**45.7**
Train	67.4	**68.5**	**77.8**	**77.8**	76.2
TV/monitor	50.9	**54.5**	61.2	64.7	**64.8**
mean AP	46.3	**47.9**	54.9	N/A	**56.1**

Table 1. Category and average precisions of the winner's method and the proposed method on the **validation set** and the **test set** of PASCAL 2008.

dissimilarities, the results on PASCAL 2008 clearly demonstrate the superiority of the proposed methods.

4.3. PASCAL VOC 2005

We further test our method on PASCAL VOC 2005 dataset [8]. This dataset contains both classification and detection tasks. The classification task contains a easy test set (test1) and a difficult test set (test2). We focus on the difficult set because the performance on the easy set is saturated. The set contains four categories (motorbike, bicycle, car and persons) and 1543 images (the images including different objects are counted only once). The best score in the competition of test2 is achieved in [8] using χ^2 distance with an extended Gaussian kernel. After that, a similar result is achieved in [37] using the EMD distance. A higher score is presented in [19] using Proximity Distribution kernel (PDK), which takes the geometric context information into consideration.

We follow the experimental setup in [8]. Specifically, we use Harris-Laplace detector and SIFT descriptor and 1000 visual words by k-means from training features. For fair comparison, we use standard codebook instead of the kernel codebook. The main difference between their method and proposed one is that we use L_1-BRD instead of χ^2. The same extended Gaussian kernel is used to incorporate

	Motor	Bike	Person	Car	Average
Winner (χ^2) [8]	**79.8**	72.8	71.9	72.0	74.1
Winner (EMD) [37]	79.7	68.1	**75.3**	74.1	74.3
PDK [19]	76.9	70.1	72.5	**78.4**	74.5
L_1-BRD	79.1	**75.4**	73.9	78.2	**76.7**

Table 2. Correct classification rates (at equal error rates) on the PASCAL challenge 2005 Test Set 2.

Figure 5. Example images from Oxford Flowers, one per class.

L_1-BRD in SVM classifier. Libsvm [4] is used and the parameter of SVM is decided by two-fold cross-validation on training set.

Table 2 compares proposed results with several state-of-the-art. We can see that we get the best average equal error rate and improve previous result by 2.2%. Although we only get one best result out of four categories, the results of other categories are also comparable to the best. This implies that the proposed L_1-BRD, which combines L_1 distance and ratio-bin dissimilarity, is relatively stable over different categories.

4.4. Oxford Flowers

The Oxford Flowers dataset [22] (see Fig. 5 for examples) contains images from 17 flower categories with 80 images per category. For each category, 40 images are used for training, 20 for validation and 20 for testing. For comparison, we directly use the three splits of this dataset provided by the authors of [22], then run three times with these splits and report the average accuracy and variance. We use the same experimental setup as for PASCAL 2008 except that standard SVM is used instead of SRKDA. The configuration include: two types of sampling – Harris-Laplace and dense sampling; five types of descriptors – SIFT and four types of color-SIFT descriptors; three types of pyramids – 1x1, 2x2 and 1x3. In total, 30 (2x5x3) channels of features are used and combined by averaging the histogram dissimilarities of each channel. For each type of descriptor, a kernel codebook of 4000 codewords is used. For comparison, we consider two histogram distances: L_1-BRD and χ^2. All the experimental setups of these two distances are exactly same to avoid bias.

Table 3 summarizes the recognition accuracies of sev-

Methods	Recognition rate
Nilsback and Zisserman [22]	71.76±1.76
Varma and Ray [33]	82.55±0.34
Nilsback and Zisserman [23]	88.33±0.3
χ^2	87.45±1.13
L_1-BRD	**89.02±0.60**

Table 3. The average of per-class recognition rates of different methods on the Oxford flower dataset.

eral published methods along with the proposed methods. We see again that L_1-BRD achieves the best performance. Also, by comparing with χ^2 in the same experimental setup, the superiority of using bin-ratio information is confirmed.

5. Conclusion

In this paper we introduce a group of histogram dissimilarities called bin-ratio dissimilarities (BRD, L_1-BRD, χ^2-BRD), which are robust to normalization and partial matching problems. In addition, these BRDs capture high-order statistics between histogram bins. We experimentally compare BRDs with several state-of-the-art dissimilarity measures on four widely tested datasets for visual category classification. The proposed methods, especially L_1-BRD, generate very promising results in all experiments. In the future, we plan to investigate further the relation of BRD to other dissimilarities and we also plan to extend BRDs to more general tasks.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures - multilevel local coding for visual recognition. *INRIA research report RR-5655*, 2006. 2

[2] A. Berg, T. Berg, and J. Malik. Shape Matching and Object Recognition Using Low Distortion Correspondences. *CVPR*, 2005. 2, 3

[3] D. Cai, X. He and J. Han. Efficient kernel discriminant analysis via spectral regression. *ICDM*, 2007. 5

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2001. 5, 6

[5] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 1999. 4

[6] O. Duchenne, F. Bach, I. Kweon and J. Ponce. A tensor-based algorithm for high-order graph matching. *CVPR*, 2009. 3

[7] M. Everingham, L. van Gool, C.K.I. Williams, J. Winn and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 Results. 2008. 4, 5, 6

[8] M. Everingham, A. Zisserman, C. Williams, L. van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers and G. Dorko and the others. PASCAL 2005, 2006. 1, 2, 4, 5, 6

[9] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005. 1, 2, 4

[10] P. Gehler and S. Nowozin. On Feature Combination for Multiclass Object Classification. *ICCV*, 2009. 4

[11] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders and J.M. Geusebroek. Visual word ambiguity. *PAMI*, 2009. 2, 4, 5

[12] J. Hafner, H. Sawhney, W. Equitz, M. Flickner and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *PAMI*, 1995. 2

[13] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 2004. 4

[14] G. Lang and P. Seitz. Robust Classification of Arbitrary Object Classes Based on Hierarchical Spatial Feature-Matching. *Machine Vision and Applications*, 10(3):123 - 135, 1997. 2

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006. 2, 4

[16] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. *ICCV*, 2005. 3

[17] J. Li, W. Wu, T. Wang and Y. Zhang. One step beyond histograms: Image representation using Markov stationary features. *CVPR*, 2008. 2, 3

[18] H. Ling and K. Okada. Diffusion distance for histogram comparison. *CVPR*, 2006. 2

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

[20] S. Maji, A.C. Berg and J. Malik. Classification using Intersection Kernel Support Vector Machines is Efficient. *CVPR*, 2008. 2

[21] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, *Visual Recognition Challange Wkshp.*, 2007. 1, 2, 5

[22] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. *CVPR*, 2006. 2, 4, 6, 7

[23] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. *ICVGIP*, 2008. 2, 7

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 4

[25] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, (Eds.) Toward Category-Level Object Recognition. *PAMI*, 2007. 2

[26] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. *CVPR*, 1997. 2

[27] Y. Rubner, C. Tomasi, and L.J. Guibas. The Earth Mover's distance as a metric for image retrieval. *IJCV*, 2000. 1, 2

[28] S. Savarese, J. M. Winn, and A. Criminisi. Discriminative Object Class Models of Appearance and Shape by Correlatons. *CVPR*, 2006. 2

[29] K.E.A. van de Sande, T. Gevers and C.G.M. Snoek. Evaluation of color descriptors for object and scene recognition. *PAMI*, 2010. 5

[30] C. Schmid. Weakly supervised learning of visual models and its application to content-based retrieval. *IJCV*, 2004. 2

[31] M. Swain and D. Ballard. Color indexing. *IJCV*, 1991. 1, 2

[32] M. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K.E.A. van de Sande & T. Gevers. Visual category recognition using spectral regression and kernel discriminant analysis. *Wkshp. on Subspace*, 2009. 1, 2, 5, 6

[33] M. Varma and D. Ray. Learning the discriminative power invariance trade-off. *ICCV*, 2007. 2, 7

[34] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. "Categorizing Nine Visual Classes using Local Appearance Descriptors", *Wkshp. Learning for Adaptable Visual Systems*, 2004. 1, 2

[35] Z. Wu, Q. Ke, M. Isard and J. Sun. Bundling features for large-scale partial-duplicate web image search. *CVPR*, 2009. 2

[36] J. Wu and J.M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. *ICCV*, 2009. 2

[37] J. Zhang, M. Marszalek, S. Lazebnik and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007. 1, 2, 4, 6