Pattern Recognition 48 (2015) 894–906

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Shape classification using invariant features and contextual information in the bag-of-words model

Bharath Ramesh, Cheng Xiang*, Tong Heng Lee

Department of Electrical and Computer Engineering, National University of Singapore, Singapore, 117576

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form
15 July 2014
Accepted 17 September 2014
Available online 30 September 2014

Keywords:
Shape classification
Bag-of-words
Log-polar transform
Contextual information
Codebook selection
Clustering evaluation
Entropy

ABSTRACT

In this paper, we describe a classification framework for binary shapes that have scale, rotation and strong viewpoint variations. To this end, we develop several novel techniques. First, we employ the spectral magnitude of log-polar transform as a local feature in the bag-of-words model. Second, we incorporate contextual information in the bag-of-words model using a novel method to extract bi-grams from the spatial co-occurrence matrix. Third, a novel metric termed 'weighted gain ratio' is proposed to select a suitable codebook size in the bag-of-words model. The proposed metric is generic, and hence it can be used for any clustering quality evaluation task. Fourth, a joint learning framework is proposed to learn features in a data-driven manner, and thus avoid manual fine-tuning of the model parameters. We test our shape classification system on the animal shapes dataset and significantly outperform state-of-the-art methods in the literature.

1. Introduction

Accurate object recognition by humans takes place due to several visual cues, such as shape, texture, color, and 3-D pose. Among these, shape is the most widely studied cue for object recognition from single two-dimensional images. Over the decades, dozens of feature descriptors have been engineered for shape analysis and classification [1–13]. In all these methods, shapes are represented either globally or locally [14]. Global approaches create a holistic representation of the shape, and so are susceptible to corruption when there is a considerable viewpoint change. On the other hand, local shape descriptors employed structurally, such as shape context [15], have been shown to be robust to deformations. This motivates us to employ the classic log-polar transform (LPT) [16] as a local descriptor, which converts scale and rotation changes in the image domain to horizontal and vertical translations in the log-polar domain, respectively. Therefore, by obtaining the Fourier transform modulus of the log-polar sampling, scale and rotation invariance can be enforced. Note that we consider the classic LPT, that is, sampling the image at the intersection of rings and wedges instead of log-polar histograms used in shape context. Since LPT is proposed as a local shape descriptor, the bag-of-words model is one of the promising choices for performing classification.

The bag-of-words model has recently emerged as the dominant framework in image classification tasks, such as object and scene classification [17–20]. First, keypoint detection [17,21] or dense sampling [22,23] is done on the image to select patches of interest, followed by a description of each patch using SIFT [17,24], raw patch [21,25] or filter-based representations [22,26]. Subsequently, the descriptors are quantized using a visual vocabulary that is commonly built using K-means [22,17]. Finally, the histograms of the training images are used to train a linear/non-linear classifier. The bag-of-words framework was applied to shape classification with some success [27], which motivates us to employ it in this work.

The major disadvantage of the bag-of-words framework is the lack of spatial information in the histogram representation. This problem was alleviated by the introduction of spatial pyramid matching (SPM), which divides the image into increasingly finer regions and constructs a histogram for each region [28]. This results in a histogram representation with a dimension equal to the number of regions times the codebook size. Spatial pyramid matching has been widely applied to scene classification tasks and it is also responsible for inspiring an array of works for the feature pooling step [29–32]. In general, higher classification accuracy has been linked to a larger vocabulary [28,33], but saturation can be expected at some point [33]. In light of this fact, the histogram obtained from the SPM approach using a large codebook is very high-dimensional (21 times the codebook size for the standard 1×1, 2×2 and 4×4 representation), which compromises on

* Corresponding author. Tel.: +65 6516 6210; fax: +65 6779 1103.
E-mail address: elexc@nus.edu.sg (C. Xiang).

http://dx.doi.org/10.1016/j.patcog.2014.09.019
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

training time and classification accuracy due to the 'curse of dimensionality' problem [29].

The Markov stationary features (MSF), first proposed in [34], provide an interesting alternative for encoding spatial information by using the spatial co-occurrence matrix [35]. To obtain the Markov stationary features, the stationary distribution of the corresponding transition matrix is concatenated with the approximate auto-correlogram features. Although the stationary distribution is a unique method to extract features, it requires calculation of higher powers of the transition matrix (typically 50) which can be extremely prohibitive for large codebooks. Moreover, the stationary distribution is an indirect method to capture information from the spatial co-occurrence matrix. In order to find an intuitive, yet a computationally less intensive way to encode contextual information, we consider the image as an article written using many "visual" words in the bag-of-words framework. Therefore, the problem of image processing is similar to language processing. In the domain of natural language processing (NLP) [36], which gave birth to the bag-of-words representation, contextual information is commonly incorporated using the N-gram model for text classification. Inspired by this idea, we interpret each entry in the spatial co-occurrence matrix as a bi-gram count. Although interpreting the spatial co-occurrence matrix as a bi-gram count is not a new idea [37], we propose a novel method to extract bi-grams using the corresponding transition matrix. For extracting the most discriminative bi-grams while reducing computational load, the training data is used for mining frequently occurring bi-grams that appear within the same shape category. Besides improving the histogram representation in the bag-of-words model, choosing the codebook size and selection of local feature parameters also play a vital role in obtaining high classification rates. The following paragraph discusses these issues.

There are two very important considerations while using the bag-of-words model: the extracted features of the image and the size of the codebook. Most methods in the literature use a codebook deemed to be large enough, simply by trial-and-error, without using a solid criteria. However, there are a handful of recent works in the literature [38–41] addressing the problem of codebook size selection. In Ref. [40], an iterative method was designed for obtaining a codebook by merging two clusters that have minimum loss of mutual information. The input to the iterative method is a codebook generated by K-means, and thus inconveniently requires selecting a 'good' size in the first place. Recently, Ref. [39] reformulated codebook generation in a supervised setting as a neural network model. Note that the focus of this paper is limited to unsupervised codebook generation in the traditional bag-of-words framework. In Ref. [38], conditional entropy and purity were proposed to evaluate the quality of the generated codebook. However, both these measures suffer from over-fitting, and therefore prefer arbitrarily large codebook sizes. As the number of clusters increases, purity and entropy reach their ideal values at the cost of having each sample as a cluster. A similar problem was encountered in the training of decision trees and gain ratio [42] was subsequently introduced for selecting an optimal attribute. We take inspiration from gain ratio and propose a metric for choosing an appropriate codebook size in the bag-of-words model. Additionally, we propose an iterative method to jointly tune the codebook size and the local feature parameters using the training data.

Briefly, the main contributions of the paper are as follows.

1. A novel method is proposed to construct histograms of bi-grams using the spatial co-occurrence matrix. The final histogram representation is low-dimensional (approx. seven times the codebook size) while significantly outperforming the original bag-of-words model and SPM [28].
2. A novel metric termed 'weighted gain ratio' is proposed to select an appropriate codebook size in the bag-of-words model. The scope of this metric is not limited to the bag-of-words framework, and so it can be used for any clustering quality evaluation task.
3. The use of the classic log-polar transform as a local feature to achieve scale and rotation invariance for shape classification. In comparison to the bag-of-words model using popular feature descriptors, LPT-based bag-of-words is shown to be superior in terms of classification accuracy.
4. A joint learning framework is proposed for codebook size selection and feature learning. It is an iterative algorithm that estimates the necessary parameters: codebook size and the maximum radius of the log-polar transform. This procedure is shown to improve the classification accuracy without the need for manual fine-tuning of model parameters.

The rest of the paper is organized as follows. Section 2 presents the shape classification system with implementation details; Section 3 presents the experimental results and discussion, followed by conclusions and future work in Section 4.

2. Contextual bag-of-words model

Binary shapes are classified in a bag-of-words framework consisting of four main stages: keypoint detection, feature extraction, vector quantization, and classification. In this work, keypoint detection is simply the selection of boundary points of the binary shape. Feature extraction involves sampling the binary shape at the keypoints, using log-polar transform, followed by computing its Fourier transform modulus. For the training set, the extracted descriptors are collectively used for K-means to obtain a codebook. The quantization step is the histogram representation of each training/testing image, using the codebook generated in the previous step. Lastly, the histograms of the training images are used to train an SVM classifier. During testing, the codebook construction step is bypassed, and so a test image is simply represented using the learned codebook and classified using SVM. The block diagram of the proposed shape classification system is shown in Fig. 1.

2.1. Feature extraction

For each boundary point of the shape, log-polar sampling is accompanied by computing its Fourier transform modulus. Subsequently, the local descriptor is obtained by converting the two-dimensional Fourier transform output into a vector and performing normalization using the Euclidean norm. Note that after extracting the log-polar transform at a particular boundary point, scale and rotation invariance is enforced by obtaining its Fourier transform modulus. To the best of our knowledge, the parameter settings of LPT as a local descriptor have not been explored in the literature. So in this subsection, we present the basic implementation details of LPT, followed by its parameter settings.

2.1.1. Log-polar transform

The log-polar transform [16] simulates the foveal mechanism of the human vision system by considering an exponential sampling of the Cartesian image. In other words, there is dense sampling near the center of the log-polar grid and coarse sampling in the periphery (see Fig. 2). Let us define the mapping from Cartesian coordinates of the image – (x, y) to LPT coordinates – (ρ, θ) as follows:

$$x' = r \cos \theta, \quad y' = r \sin \theta, \tag{1}$$

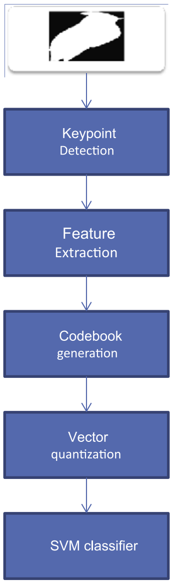


Fig. 1. Block diagram of the shape classification system.

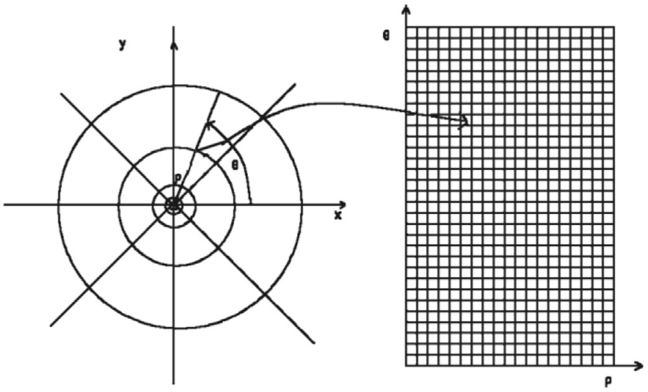


Fig. 2. Mapping from Cartesian (x, y) to log-polar space (ρ, θ). Figure modified from source: [43].

where (r, θ) are polar coordinates defined with (x_c, y_c) as the center of the transform and $(x', y') = (x - x_c, y - y_c)$, that is,

$$r = \sqrt{(x')^2 + (y')^2}. \tag{2}$$

The angle θ is required to be in the range $[0, 2\pi)$, but arctan is defined only for $(-\pi/2, \pi/2)$. Therefore, the angles are computed depending on the quadrant as shown below:

$$\theta = \begin{cases} \arctan\left(\dfrac{y'}{x'}\right) & \text{if } x' > 0 \\ \arctan\left(\dfrac{y'}{x'}\right) + \pi & \text{if } x' < 0 \\ +\dfrac{\pi}{2} & \text{if } y' > 0,\ x' = 0 \\ +\dfrac{3\pi}{2} & \text{if } y' < 0,\ x' = 0 \\ \text{undefined} & \text{if } y' = 0,\ x' = 0 \end{cases} \tag{3}$$

The above operation produces output in the range $(-\pi/2, 3\pi/2]$, which can be mapped to $[0, 2\pi)$ by adding 2π to negative values. The convention of the log-polar parameters in Young [44] has been adopted here. The radii of the smallest and the largest ring are represented as r_{min} and r_{max}, respectively. The logarithmic scaling is defined as $\rho = \log r$. The samples of LPT lie at the intersection between rings and wedges, and thus the size of the log-polar image is n_r by n_w, where n_r and n_w are the number of rings and wedges, respectively. In general, the intersection happens at arbitrary locations in the image, and therefore bilinear interpolation is used to find the image intensity at these locations. Bilinear interpolation considers the closest 2×2 neighborhood of known pixel values surrounding the unknown value. For an image I, if (x, y) is the location of the unknown value and (x_1, y_1), (x_1, y_2), (x_2, y_1) and (x_2, y_2) are the surrounding pixel locations, then the image intensity $I(x, y)$ is given by a weighted summation:

$$I(x,y) = \frac{(x_2 - x)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} I(x_1, y_1) + \frac{(x - x_1)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} I(x_2, y_1) + \frac{(x_2 - x)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)} I(x_1, y_2) + \frac{(x - x_1)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)} I(x_2, y_2). \tag{4}$$

2.1.2. *Setting of LPT parameters*

Scale and rotation changes in the Cartesian image correspond to horizontal and vertical shifts in the log-polar domain, respectively [16]. Note that the Fourier transform modulus of two images related by pure translation is the same. Consequently, two log-polar images of similar shapes, which have scale and rotation variations, are expected to have "similar" Fourier transform magnitude. This concept is illustrated in Fig. 3 for the simple case of a circle to facilitate easy visual comparison in the frequency domain. After eliminating the translation differences in the log-polar space, it is easy to see that circles of different radii have nearly identical features in the frequency domain. Therefore, our choice of LPT is motivated by this sound mathematical basis for ensuring scale and rotation invariance, which many heuristically designed feature descriptors (SIFT [45], LBP [46], etc.) fail to achieve theoretically. Recently, an attempt [47] was made to theoretically explain the phenomenal success of the SIFT descriptor. This study has proved that SIFT is scale and rotation invariant under certain conditions. However, scale and rotation invariance is achieved only for the selected keypoints using the scale-space and does not apply to other useful structures, like edges in the image [48].

On the other hand, log-polar transform is sensitive to changes in the location of its centroid on the image, i.e., greater the center mismatch between two shapes, greater the image distortion [16]. In other words, occlusion and viewpoint change would severely affect the invariant properties of LPT. To address this issue, we choose to place the centroid of LPT at the shape boundaries and extract local features instead of a global representation, as shown in Fig. 4. This line of reasoning is backed up by the dominance of local feature-based approaches (over global approaches) for various recognition tasks in computer vision [49]. Even so, we verify our choice of the local approach by comparing it with the global application of LPT.

For each training/testing image, log-polar sampling is done by centering on the shape's centroid, followed by computing the Fourier transform modulus to obtain a global shape descriptor (Fig. 3 shows this step). Then the descriptors from the training set are analyzed to find a discriminant low-dimensional subspace, using Principal Component Analysis (PCA) [50] and Recursive Fisher's Linear Discriminant (RFLD) [51]. After projecting the training descriptors to the subspace, the resultant descriptors are used to train an SVM classifier. For a test image, the Fourier transform modulus of LPT is projected to the low-dimensional subspace and classified using SVM. Table 1 compares the classification accuracies on the animal shapes database [52] using the global and local approach. It is clear that the local approach easily outperforms the global approach, which reaffirms our choice of local LPT descriptors and the bag-of-words model. Due to space constraints, we only report the result using the best parameter settings of LPT for the global approach – minimum radius $r_{min} = 1$,

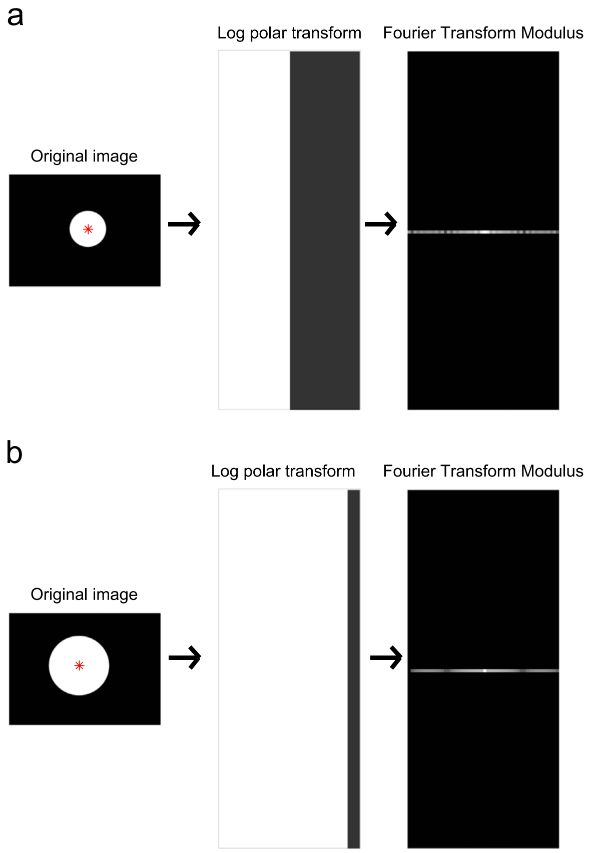


Fig. 3. (a) and (b) Log-polar transform applied to the image by centering on the shape, followed by computing the Fourier transform modulus. Scale change in the Cartesian space corresponds to a horizontal shift in the log-polar space, which can be eliminated by computing the Fourier transform modulus to obtain a scale invariant descriptor for each binary shape.

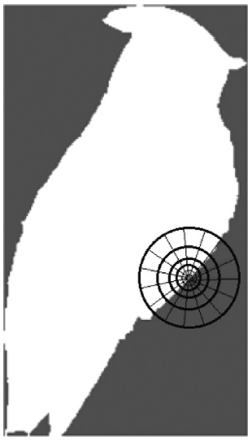


Fig. 4. Feature extraction using log-polar transform at the shape boundaries.

Table 1
Comparison of local and global approach using LPT (%).

	Local LPT approach	Global LPT approach
Accuracy	78.30	53.70

maximum radius r_{max}=max. radius of shape image, number of rings n_r=120, and the number of wedges n_w=180. The parameter settings of the proposed LPT local approach are discussed in detail below.

Apart from the centroid location, there are four other parameters for the log-polar sampling: minimum radius (r_{min}), maximum radius (r_{max}), number of rings (n_r), and number of wedges (n_w). Intuitively, the minimum radius would not play a significant role as feature extraction is done for every boundary point. Ref. [53] recommends 3–5 pixels as the minimum radius for object detection in grayscale images. However, shape context [15] obtained good results with 2 pixels for shape classification in binary images. Similarly, we found that using a minimum radius of 2 pixels is one of the best in terms of classification accuracy (Fig. 5 (a)). In the literature, shape context [15] quantized the angle into 12 divisions (n_w) and log-distance into 5 divisions (n_r). It could afford such a coarse sampling (5 × 12) due to the histogram-style treatment of log-polar transform. Nevertheless, when using the original log-polar transform, it was found that a denser sampling is required to obtain higher classification accuracies (Fig. 5(b)). Spatial pyramid matching [28] was used to obtain the classification accuracy for the comparative studies shown in Fig. 5 and for the local approach in Table 1.

It is not straightforward to determine the maximum radius of LPT as a local feature; too small a radius will make the feature non-discriminatory and too large a radius would not make sense as a local feature. Ref. [44] makes a reasonable suggestion to keep every pixel's orthogonal neighbors about equal distances from it, by applying the following constraint:

$$r_{max} = r_{min} \times e^{2\pi((n_r-1)/n_w)}. \tag{5}$$

Using $r_{min}=2$, n_r=14, and n_w=30 in Eq. (5) results in a maximum radius of about 30 pixels, which was used for illustration in Fig. 4. This may still be sub-optimal in terms of classification accuracy, and so a procedure to set the maximum radius in a data-driven manner is presented later in this section.

2.2. Codebook selection

The descriptors obtained from the training images are collectively used to obtain a codebook, using VLFeat's [54] implementation of K-means with accelerated Elkan algorithm for optimization. The codebook is simply the collection of the cluster centroids obtained using K-means. In order to evaluate the clustering quality (the discriminative power of the codebook), many measures have been proposed in the literature. Those include combinatorial techniques [55], and external cluster evaluation measures like *F*-measure [56], misclassification index (MI) [57], among others. Out of the external evaluation measures, the ones based on information theory, like purity and conditional entropy, are independent of the size of the data set, the number of clusters and the clustering algorithms used. This provides information theory based measures – a unique advantage over other classes of measures [58,59]. Following this reasoning, Ref. [38] proposed the use of purity and conditional entropy as evaluation measures for visual codebooks. However, both these measures improve with an increase in the number of clusters, up to a degenerate maximum where there are as many cluster centers as data points. Therefore, clustering evaluations based on these metrics are biased and score high on suboptimal solutions [58]. The rest of this subsection presents the details about entropy-based measures, their drawbacks, and the proposed metric.

Let $P=\{p_1, p_2, \ldots, p_C\}$ represent the probability distribution of the training descriptors belonging to C shape categories. Then the information conveyed by this distribution, entropy of *P*, is given by

$$\text{Info}(P) = -\sum_{j=1}^{C} p_j \log_2 p_j, \tag{6}$$

$$p_j = N_j/N \tag{7}$$

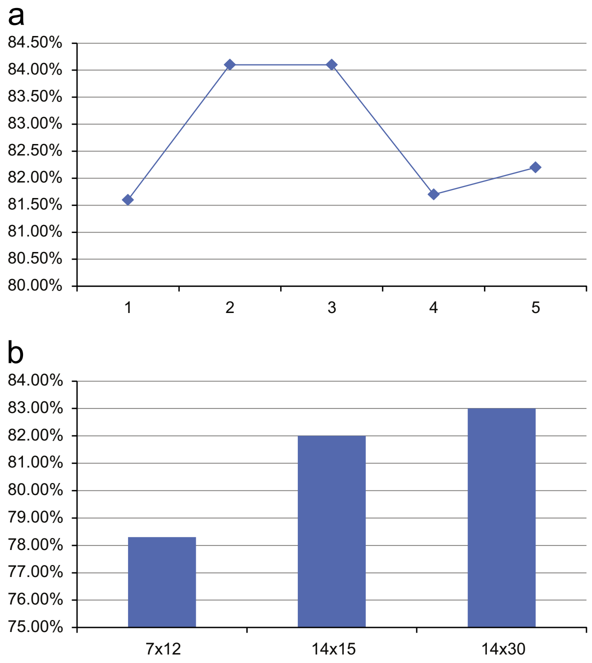


Fig. 5. (a) LPT minimum radius vs. classification accuracy. (b) LPT size vs. classification accuracy.

where N_j is the number of data points belonging to class j and $N=N_1+N_2+\cdots+N_C$ is the total number of data points. After partitioning the data into K clusters, the entropy of each cluster E_i is given by

$$E_i=-\sum_{j=1}^{C} p_{ij}\log_2 p_{ij}, \quad i=1,2,\ldots,K \tag{8}$$

where p_{ij} is the ratio of number of samples of class j in cluster i (n_{ij}) to the total number of samples in cluster i (n_i),

$$p_{ij}=n_{ij}/n_i. \tag{9}$$

The entropy of the codebook is the weighted average of the entropies of the K clusters,

$$\text{Info}(P,K)=\sum_{i=1}^{K} p_{c_i}E_i \tag{10}$$

where p_{c_i} is the ratio of the number of samples in cluster i (n_i) to the total number of samples (N),

$$p_{c_i}=n_i/N. \tag{11}$$

Thus the information gain, denoted as Gain(P, K), is defined as,

$$\text{Gain}(P,K)=\text{Info}(P)-\text{Info}(P,K). \tag{12}$$

In order to maximize information gain, the entropy of the codebook Info(P, K) is to be minimized. This quantity goes to zero in an undesirable fashion when every sample or data point is treated as a cluster. In the machine learning domain, the induction of ID3 decision trees suffered from a similar problem and was rectified by normalizing the information gain using the split information [42]. The split information takes into account the number of data points in the clusters and prevents over-fitting. By normalizing information gain by the split information of the codebook, defined in Eq. (14), the resultant term,

$$\text{GainRatio}(P,K)=\frac{\text{Info}(P)-\text{Info}(P,K)}{\text{SplitInfo}(P,K)} \tag{13}$$

can be maximized. In the context of clustering, split information is given by

$$\text{SplitInfo}(P,K)=-\sum_{i=1}^{N} p_{c_i}\log_2 p_{c_i}. \tag{14}$$

Ideally, split information should increase significantly with increase in codebook size, and in turn, leads to a decrease in gain ratio before reaching a very large codebook size. However, from experiments up to a very large codebook size of 20,000, gain ratio kept increasing without attaining a maxima, as illustrated in Fig. 6. This observation is supported by Ref. [60], which demonstrated that gain ratio is still biased in favor of attributes with large number of values. To address this problem, we propose the following metric which considers the ‘physical size’ of the clusters in the codebook.

$$\text{WeightedGainRatio}(P,K)=\frac{\text{Info}(P)-\text{Info}(P,K)}{\text{SplitInfo}(P,K)\times\text{VarianceRatio}(P,K)} \tag{15}$$

where the proposed ‘weight term’ is defined as

$$\text{VarianceRatio}(P,K)=\frac{V_K}{V_{avg}}. \tag{16}$$

The variance of the codebook V_K is defined as

$$V_K=\frac{d_{c_1}^2+d_{c_2}^2+\cdots+d_{c_K}^2}{K} \tag{17}$$

where d_{c_i} represents the Euclidean distance between the centroid of ith cluster to the mean of all centroids. When the number of clusters is small, we can expect Eq. (17) to have a small value and increase as the number of clusters increases. The variance of a single cluster in the codebook is similarly defined as

$$V_i=\frac{d_{1i}^2+d_{2i}^2+\cdots+d_{n_i i}^2}{n_i} \tag{18}$$

where d_{ki} is the Euclidean distance between the kth member of cluster i to the cluster centroid. The average variance of the clusters is obtained by taking the mean value of all cluster variances:

$$V_{avg}=\frac{1}{K}\sum_{i=1}^{K} V_i. \tag{19}$$

As the codebook size increases, the average cluster size is expected to decrease. Therefore, the weight term (Eq. 16) increases in magnitude as the codebook size increases, as shown in Fig. 7(a). Notice that ‘weighted gain ratio’ leverages on both the number of

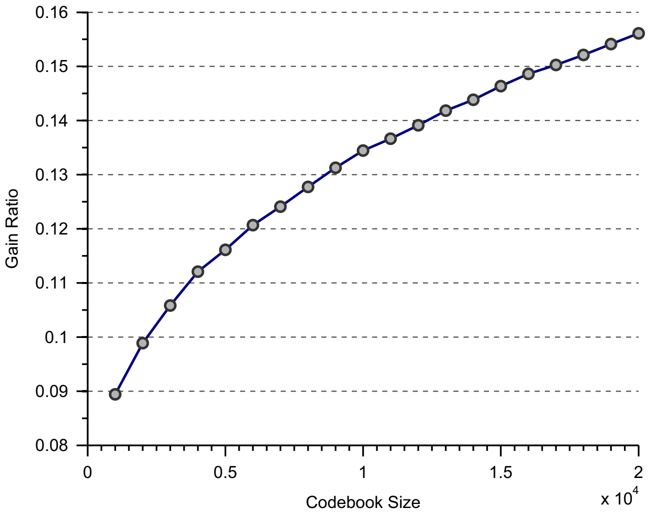


Fig. 6. Gain ratio for codebook sizes up to 20,000.

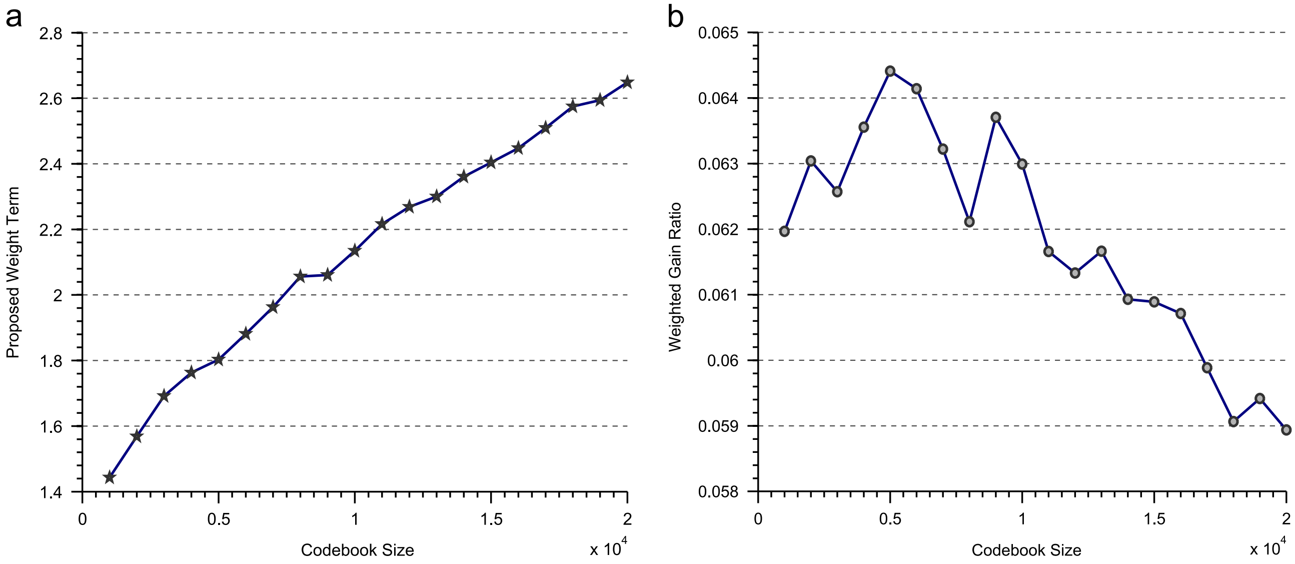


Fig. 7. Sample trends of the proposed weight term and metric for codebooks corresponding to Fig. 6. (a) Codebook size vs. proposed weight term. (b) Codebook size vs. weighted gain ratio.

data points in a cluster (split information) and the size of each cluster (variance term) to account for the change in codebook size. Since it is possible to have many points in a cluster and still have a small cluster size, and vice versa, it is important to use both split information and the variance term together. In other words, the proposed weight term is expected to increase the rate of change of split information, and thus avoid very high codebook sizes. So, if we have a set of codebooks, it is possible to select one that maximizes 'weighted gain ratio', as illustrated in Fig. 7(b).

2.3. Joint learning framework

If the extracted local features of the image are not discriminatory, then optimizing the codebook size is meaningless. So, we formulate an iterative approach to feature learning and codebook size selection (refer to Algorithm 1). Using the initial LPT parameter settings, the first iterative step of the joint learning framework selects the codebook with maximum weighted gain ratio (Eq. (15)). For the obtained codebook size, the second step selects a codebook which maximizes gain ratio (Eq. (13)) among codebooks with different values of LPT r_{max}. These two steps are iterated until there is convergence of codebook size and LPT maximum radius. Note that 'weighted gain ratio' is not used for the second iterative step, because of fixing the codebook size from the output of the first step. The output of the joint learning framework is a codebook of a particular size and LPT r_{max}, which is then used to represent the training/testing images. In the next subsection, details about the proposed histogram representation are presented.

Algorithm 1. Joint learning algorithm.

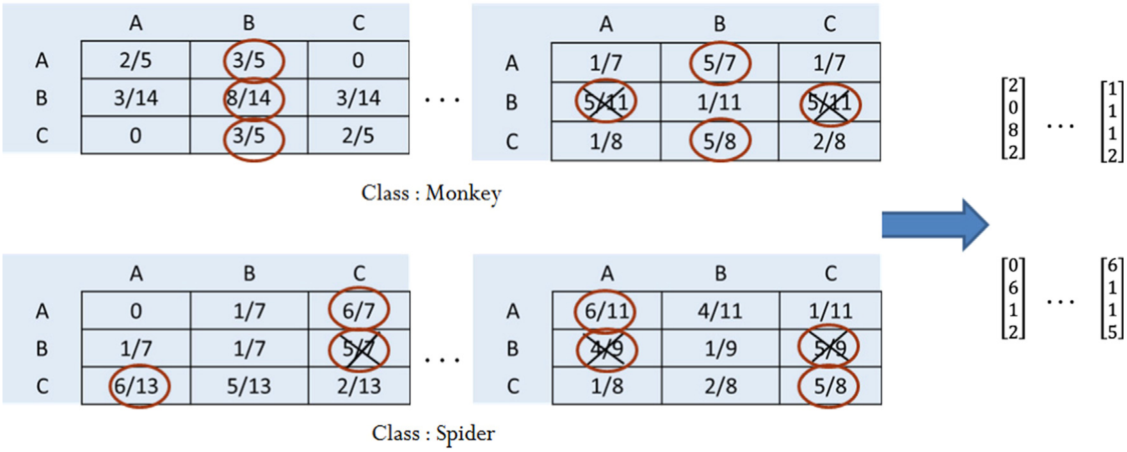
```
Set the feature parameters $n_r = 14, n_w = 30, r_{min} = 2$ and
initialize $r_{max}$ using Eq. (5).
repeat
    Choose a codebook with size $K \in (1000, 2000, \ldots 20,000)$
    using weighted gain ratio.
    Fix codebook size $K$ from previous step.
    Choose a codebook with LPT $r_{max} \in (5, 10, \ldots, 125)$ using
    gain ratio.
    Fix LPT $r_{max}$ from previous step.
until no change in $K$, $r_{max}$
Output: Codebook of size $K$ using LPT features with $r_{max}$.
```

2.4. Contextual information

The bag-of-words histogram representation discards the spatial relationship between local features, and hence cannot differentiate between local features with different spatial configurations. As mentioned earlier in Section 1, spatial pyramid matching [28], which has been very successful for scene classification, was proposed to encode spatial relationships between the local features. However, due to its high-dimensional histogram representation, some works have opted for compact representations [29], or Markov stationary features [34], or higher order spatial co-occurrence statistics [61]. Higher-order statistics can yield richer information, but in applications involving sparse sampling of the image (approx. 3% of the total image pixels in this work), it may not be readily derivable. On the other hand, the Markov stationary features (MSF) proposed in [34] provide an attractive alternative for encoding spatial information using the spatial co-occurrence matrix [35]. However, the stationary distribution of MSF is a computationally intensive and an indirect way to capture information from the spatial co-occurrence matrix. Notice that the problem of image processing is similar to language processing, because the image can be considered as an article written using many "visual" words of the codebook. Inspired by this idea, we interpret each entry in the spatial co-occurrence matrix as the pairwise occurrence count of the codewords, or in other words, bi-gram count. The following paragraphs present the details about the proposed bi-gram extraction procedure.

For each training/testing image, feature extraction followed by vector quantization enables each local descriptor to be assigned to one of the K visual words. So, let us define an image I_{ind}, having the word indices – $\{1, 2, \cdots, K\}$ – as pixel values. The word indices simply represent the K visual words, $S = \{c_1, c_2, \ldots c_K\}$, assigned to the LPT descriptors during vector quantization (Section 2.5). In this work, the boundaries of the binary shape are assigned to a particular index of the visual word and other locations, where local features are not extracted, are set to zero. Therefore, each pixel $I_{ind}(x, y)$ of an m by n index image takes one of the values in the set $\{0, 1, 2, \ldots, K\}$. The spatial co-occurrence matrix is created by calculating how often a pixel value i $(i \neq 0)$ occurs adjacent to a pixel with the value j $(j \neq 0)$. We denote the co-occurrence matrix as $C \in R^{K \times K}$, in which each entry is computed as follows:

$$C(i,j) = \sum_{x=1}^{n} \sum_{y=1}^{m} \#(I_{ind}(x,y) = i, I_{ind}(x_n, y_n) = j) \tag{20}$$



	A	B	C
A	2/5	3/5	0
B	3/14	8/14	3/14
C	0	3/5	2/5

	A	B	C
A	1/7	5/7	1/7
B	~~5/11~~	1/11	~~5/11~~
C	1/8	5/8	2/8

Class : Monkey

	A	B	C
A	0	1/7	6/7
B	1/7	1/7	~~5/7~~
C	6/13	5/13	2/13

	A	B	C
A	6/11	4/11	1/11
B	~~4/9~~	1/9	~~5/9~~
C	1/8	2/8	5/8

Class : Spider

Fig. 8. An example of the bi-grams extraction procedure.

where every pixel location (x,y) and its immediate neighbors (x_n,y_n) satisfying $0<\sqrt{(x_n-x)^2+(y_n-y)^2}\le\sqrt{2}$ are inspected to count the number of $i-j$ pairs. Simply put, each entry $C(i,j)$ in the spatial co-occurrence matrix records the total number of times a pair of neighboring local descriptors get assigned to c_i and c_j, which are any two of the K visual words denoted as $S=\{c_1,c_2,\dots c_K\}$.

The corresponding transition matrix $T\in R^{K\times K}$ is defined as

$$T(i,j)=\frac{C(i,j)}{\sum_{k=1}^{K}C(i,k)}. \tag{21}$$

Since the spatial co-occurrence matrix is row normalized to obtain the transition matrix (Eq. (21)), each element of the transition matrix represents the pairwise occurrence probability of the codewords, or in other words, bi-gram occurrence probability. So to extract discriminatory features, we select bi-grams with high probability from all the transition matrices of the training data, and subsequently discard those that occur across different shape categories. As a result, it is possible to retain a unique signature for each category in the histogram representation, and at the same time reduce computational load. The selected bi-grams are termed as ‘class-unique bi-grams’, because of appearing with high probability within training images of a single shape category. After selecting the class-unique bi-grams, the spatial co-occurrence matrix characterizes their frequency for each training/testing image. This procedure is described below using an example.

The procedure for extracting class-unique bi-grams is illustrated in Fig. 8, in which three sample codewords denoted as A, B, and C are used to represent the training images from two classes as 3 by 3 transition matrices (refer to [34] for a visual description of obtaining the transition matrix from the image). Notice that the spatial co-occurrence matrix can be easily derived from the transition matrices in Fig. 8, by simply considering the numerator term in each entry. For instance, the AA bi-gram in the first matrix of the monkey class has occurred twice, the AB bi-gram thrice, etc. So in total, the codeword A has been assigned five times to a local descriptor, whose neighboring descriptors have been assigned to either A itself or the codeword B. By setting a threshold of 0.4 for each transition matrix, the circled entries represent those bi-grams which are above this probability threshold. The potentially confusing bi-grams (BA and BC), which appear in both the classes with a high probability, are discarded and the remaining seven entries – AA, AB, AC, BB, CA, CB, CC – are further investigated. Since the spatial co-occurrence matrix is symmetrical, duplicate entries like AC and CA are singled out and either of them are kept. In addition, entries AB and CB are removed because their symmetrical counterparts BA and BC were discarded. Finally, a 4-dimensional histogram of bi-grams using AA, AC, BB and CC is created using the corresponding spatial co-occurrence matrix. Note that it is easy to implement this algorithm using the MATLAB commands – *fliplr* (check symmetry of the matrix indices) and *unique* (extract class unique bi-grams and remove duplicate entries).

2.5. Vector quantization and classification

A training/testing image is quantized into K histogram bins, i.e., the local features extracted from an image are matched to the nearest visual word using Euclidean distance, one by one, and the frequency of each word creates the K-dimensional histogram representation. Let the number of class-unique bi-grams be N_{bi}. The normalized bag-of-words representation is concatenated with the normalized histogram of bi-grams, to produce a vector of dimension – $(K+N_{bi})$. Besides the bi-gram features, a 2×2 image grid is used to capture mid-level spatial information. Each of the four histograms from the 2×2 grid is normalized separately and concatenated together. In turn, the $4K$-dimensional vector using the 2×2 grid is concatenated with the $(K+N_{bi})$-dimensional vector to form the final $(5K+N_{bi})$-dimensional representation for each image. The classifier used is the SVM implementation of VLFeat in their bag-of-words application [54].

3. Experiments and results

We test our shape classification system on the animal shapes database, which was introduced by [52], consisting of 2000 binary shapes of 20 animal categories with 100 shapes for each category. It is a challenging dataset with very large intra-class variations, strong inter-class similarities among some categories, viewpoint changes and occlusion (Fig. 9). The database was randomly split into half for training and half for testing, following the protocol in [52,62,27,63]. The experiments were run on HP Xeon Two Sockets Quad-Core 64-bit Linux clusters with 200 GB memory limit.

3.1. Joint learning results

The steps taken by the joint learning framework are shown in Fig. 10. After LPT r_{max} was initialized using Eq. (5), weighted gain ratio peaked at a codebook size of 7000 in the first iteration (Fig. 10(a)). During the next phase to select r_{max}, gain ratio showed an increasing trend as expected, and eventually saturated for codebooks with a very high LPT r_{max} (Fig. 10(b)). Therefore, a relative change threshold of 1% was used for selecting a codebook

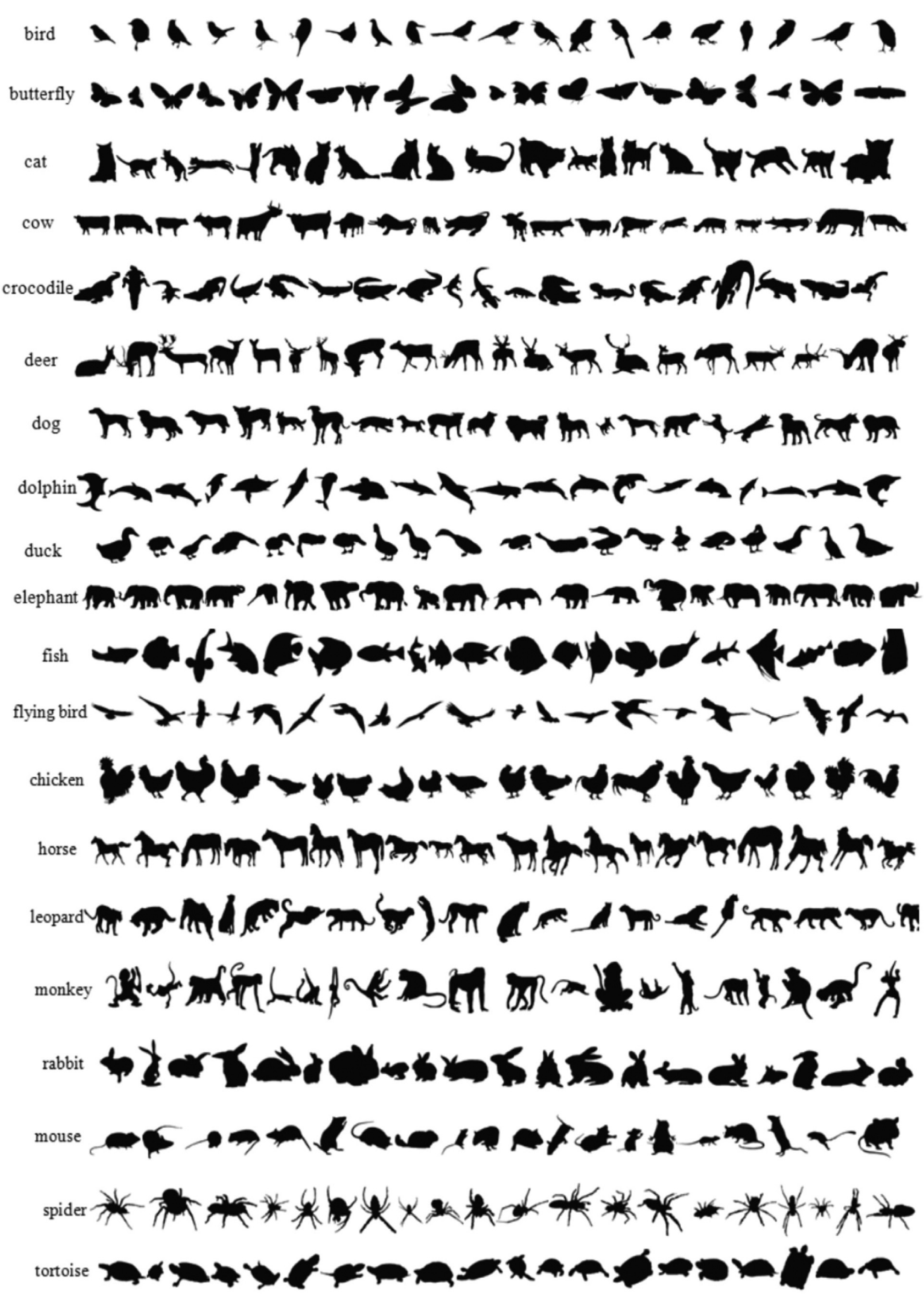


Fig. 9. Samples from the animal shapes dataset. Figure adapted from the website of the first author in Ref. [52].

with a moderate LPT $r_{max} = 75$, which corresponds to the codebook with the maximum gain ratio before reaching the threshold. After setting $r_{max} = 75$, weighted gain ratio peaked at a lower codebook size of 5000 in the next iteration (Fig. 10(c)). This is possibly due to the selection of a better LPT parameter from the first iteration. In the next phase, the same maximum radius was selected based on a relative change threshold of 1% for (Fig. 10(d)). Therefore, a couple of iterations were sufficient to converge on the final codebook size to be 5000 with LPT $r_{max} = 75$. Next, we investigate whether the parameters selected by the joint learning algorithm give high classification accuracy compared to a wide range of settings.

Fig. 11 shows the classification accuracy obtained with different codebook sizes and LPT r_{max}. For the bag-of-words model, codebooks with LPT $r_{max} > 40$ give higher classification accuracy compared to the codebook with $r_{max} = 30$ initialized using Eq. (5) (refer to Fig. 11(a)). Although the codebook with $r_{max} = 75$ did not give the highest accuracy, the effectiveness of the joint learning strategy is clearly evident. A similar trend can be observed for the spatial pyramid approach using codebooks with different maximum radius of the log-polar transform. For the proposed method, the selected codebook with LPT $r_{max} = 75$ provides a 2% boost in classification accuracy compared to the codebook initialized with LPT $r_{max} = 30$ (Fig. 11(a)). By inspecting Fig. 11(b), it is clear that the codebook size selected using the joint learning algorithm (5000) gives a high classification accuracy, which is close to the highest obtained for the bag-of-words model. Larger codebooks may provide higher classification accuracy for the bag-of-words representation, but it can drop considerably as seen from the trend towards a codebook size of 20,000. However, SPM provides good classification accuracy for codebook sizes only up to 6000, and suffers from the high-dimensionality of the histogram (21 K) for higher codebook sizes (refer to Fig. 11(b)). In comparison to SPM, the incorporation of bi-gram features in the bag-of-words model provides higher classification accuracy, and also gives a relatively stable accuracy for codebook sizes up to 10,000 (Fig. 11(b)). In summary, we have shown that there is a positive correlation between the model parameters selected using the joint learning algorithm and the classification accuracy.

3.2. Classification results

To demonstrate the effectiveness of the proposed histogram representation, the performance of our method is compared with four baseline methods in Table 2, where "Bi-gram" is the

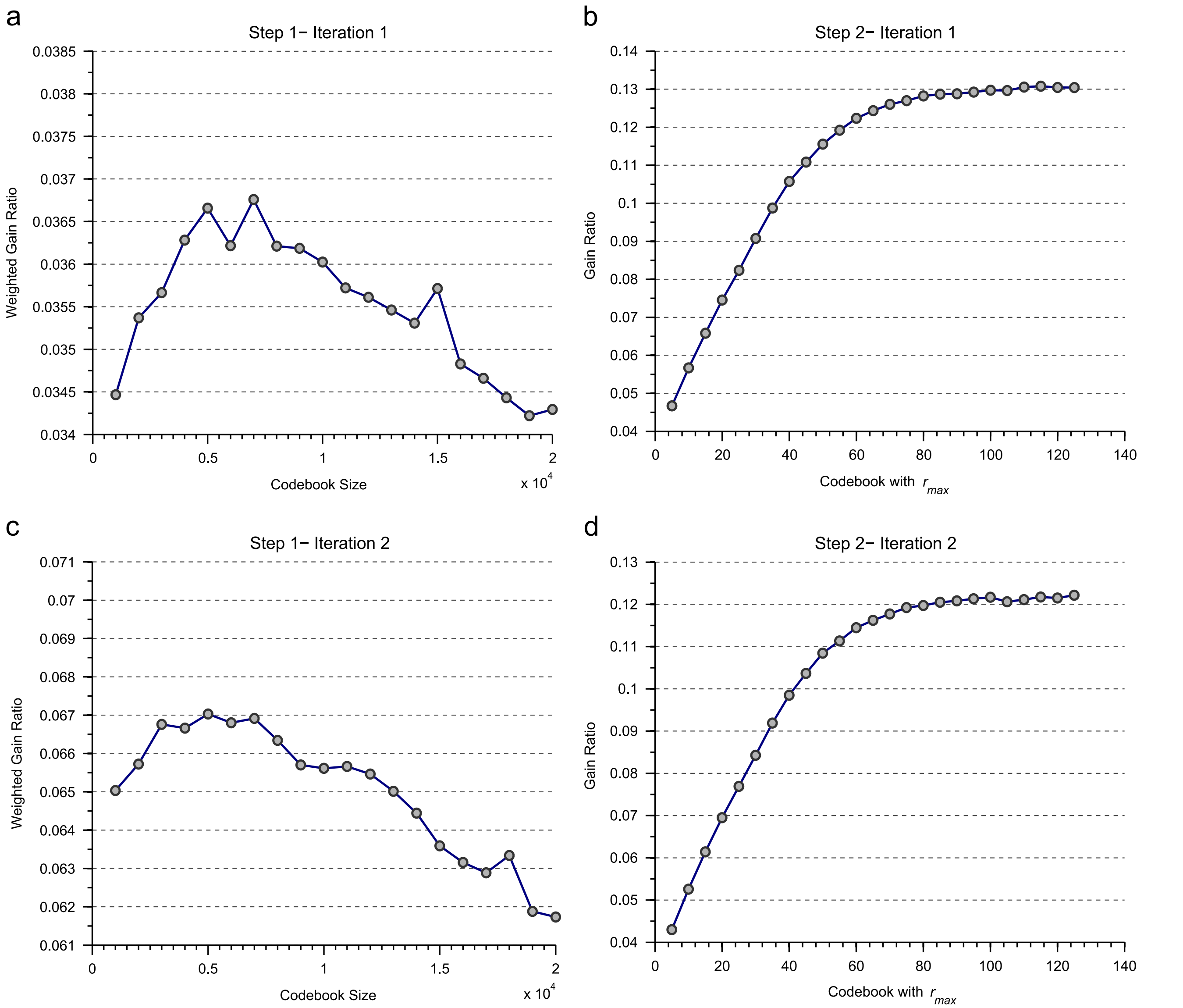


Fig. 10. Results from the joint learning framework. (a) LPT r_{max} initialized using Eq. (5). (b) Codebook size set to 7000. (c) LPT r_{max} set to 75. (d) Codebook size set to 5000.

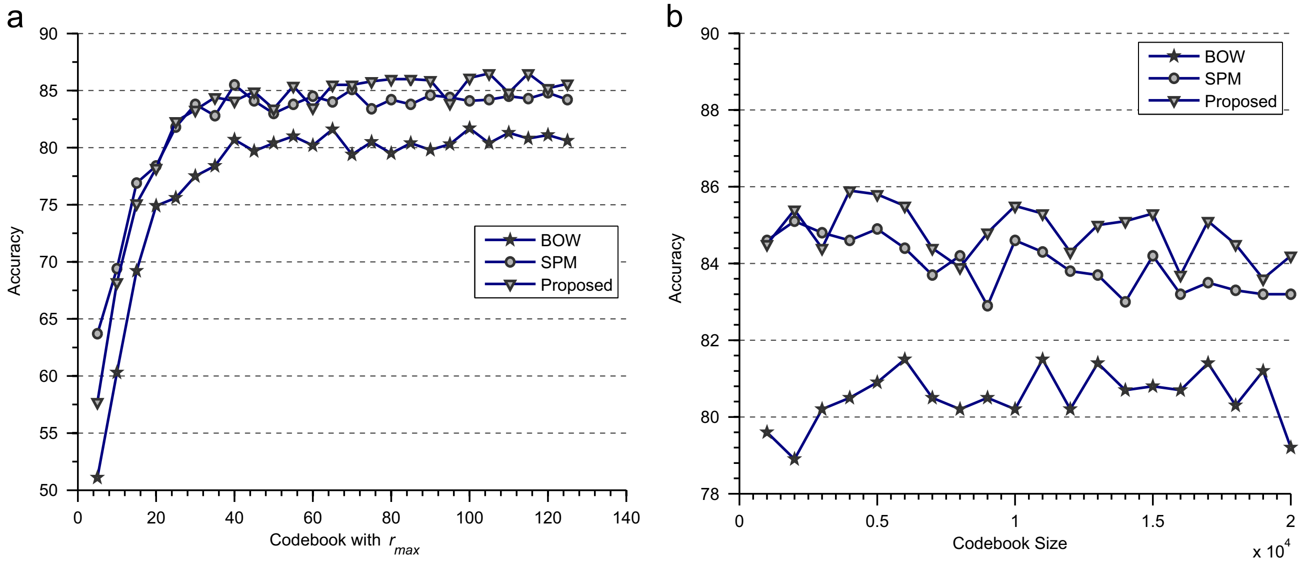


Fig. 11. Classification accuracy obtained with three methods for different LPT maximum radius and codebook sizes (%). (a) Fixed codebook size = 5000. (b) Fixed LPT $r_{max} = 75$.

representation with the global BoW histogram coupled with the class-unique bi-grams, and “Proposed” is the 2×2 grid representation added to the “Bi-gram” model. Note that the Markov stationary feature (MSF) [34] performs on par with the bag-of-words model. In contrast, we show that the spatial co-occurrence matrix can be exploited in a much simpler way to boost the classification accuracy (“bi-gram”). The results reported for the proposed method are using a threshold probability of 0.7 for the transition matrix, resulting in 11,275 class unique bi-grams. Fig. 12 shows the effect of varying the probability threshold on the classification accuracy. It can be noted that the classification accuracy does not change significantly for different threshold values. This phenomenon can be attributed to the selection of frequently occurring bi-grams that appear within a single shape category of the training set.

Table 3 compares the proposed method with the previous works using the animal shapes dataset. It is clear that the proposed method significantly outperforms the state-of-the-art algorithms while still using a low-dimensional histogram representation. In comparison to SPM, which uses a 21K-dimensional histogram representation, the proposed method uses only a 7K-dimensional histogram representation. Note that the training and testing set were generated using a random database split – half for training and half for testing – as in [52,62,27,63]. So one may argue that the high classification result is possibly due to the “random” nature of training and testing. In order to further demonstrate the virtue of the proposed histogram representation and the local descriptor, ten-fold cross validation was done with a codebook size of 5000 and LPT $r_{max} = 75$. In particular, the dataset was split into 10 non-overlapping sets of equal size while maintaining the class balance in each split. Now we combine 9 of these for training and the remaining one is for testing. This process is repeated for other combinations of training and testing sets. We obtained 87.8% classification accuracy, using cross validation, and thus conclude that the high classification accuracy (86.0%) is due to a genuine improvement of the bag-of-words model.

Besides cross validation, we can manually verify whether the system makes reasonable mistakes by inspecting the confusion matrix (Fig. 13). We observe that animals with distinct visual attributes like spider, butterfly, hen, elephant, duck, deer, and horse contribute to a total error of just 2%. The potentially confusable ones, which share similar physical attributes, like dog, leopard, cat, and mouse pose a greater problem to the classification system. The lack of 3-D information about the shapes creates considerable ambiguity, even for humans, especially among four-legged animals and aquatic species (refer to Fig. 9). Therefore, we conclude that the proposed shape classification framework is capable of high performance under challenging conditions like scale, rotation and strong viewpoint changes.

Table 2
Performance comparison of the proposed methods with four baseline methods (%).

Alg.	BoW	MSF [34]	Triv. MSF [64]	SPM [28]	Bi-gram	Proposed
Acc.	81.1	81.1	81.5	84.8	83.5	86.0

Comparing Tables 2 and 3, LPT based bag-of-words (“BoW” in Table 2) slightly outperforms the bag-of-words representation using the popular feature descriptor SIFT [27]. However, the comparison may not be fair because of differences in implementation and other parameters. So in the next subsection, we describe our implementation of a SIFT based bag-of-words model and compare it with the proposed framework.

3.3. Comparison with SIFT

Ignoring scale selection, many works have found that SIFT sampling at multiple scales performs well in object classification systems using a bag-of-words framework, as noted in [33]. For establishing a fair comparison with the proposed method, we replace log-polar transform feature extraction step with SIFT descriptor at multiple scales, while all other components remain unchanged. The chosen scales are four, six, eight and ten, following the publicly available implementation of VLFeat's object classification system [54]. The codebook size was chosen to be 3000 using ‘weighted gain ratio’ from a range of codebooks sizes – 1000, 2000,…, 20,000. Finally, we obtained an accuracy of 80.2% using our implementation of SIFT-based bag-of-words framework, which is markedly similar to the accuracy of 80.4% obtained in [27]. Therefore, we conclude that the accuracy obtained using the proposed LPT-based method is significantly higher than SIFT-based bag-of-words for binary shape classification. In the following subsection, we compare the proposed LPT local feature with other well-established shape descriptors.

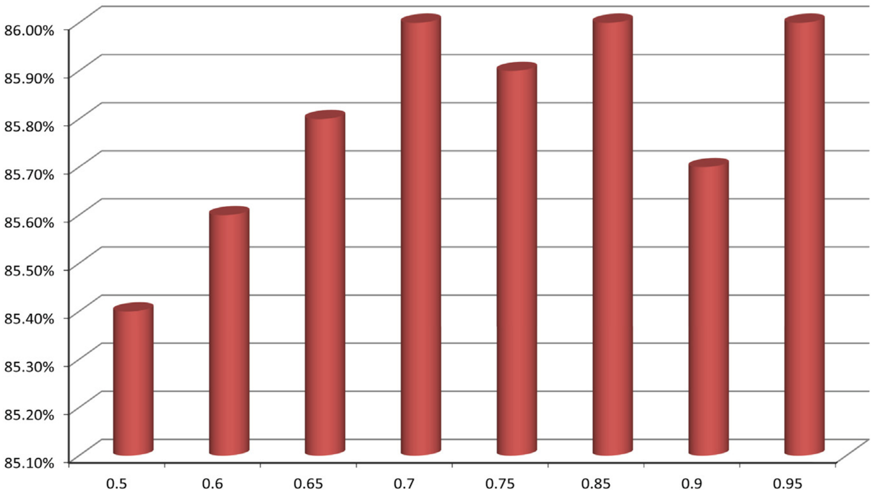


Fig. 12. Effect of changing the probability threshold for the transition matrix on classification accuracy.

3.4. Comparison with fourier descriptors

Fourier descriptors have long had a good reputation for shape representation and retrieval. Therefore, we compare two popular global shape descriptors – centroid distance signature (also known as 1-D Fourier descriptor) [67] and generic Fourier descriptor (GFD) [68] – with the global LPT approach. Note that the centroid distance signature can be readily implemented using DIPUM toolbox [69] and GFD's polar transform using the command – *cart2pol* – in MATLAB. Table 4 shows the comparison between these methods and the global LPT approach, described earlier in Section 2.1. Notice that we have not compared the Fourier descriptors with the proposed local LPT framework, because they were proposed as global shape descriptors much before the bag-of-words model became the dominant classification framework. Therefore, we extend GFD as a local descriptor and compare it to the proposed LPT local shape descriptor. It should be noted that generic Fourier descriptor is a highly related work compared to the proposed approach in this paper, because it uses LPT's counterpart – polar transform, followed by Fourier transform modulus. In other words, using GFD as a local descriptor is equivalent to replacing log-polar transform with polar transform in the proposed framework, which makes for a very interesting comparison. The parameter selection for the GFD local descriptor is explained below.

The size of the polar grid was chosen to be the same as LPT's grid size, i.e., 14 rings and 30 wedges. The minimum and maximum radius was chosen to be 2 and 40, respectively. Note that the maximum radius of the polar grid was chosen exhaustively to give the best classification accuracy. The codebook size for the GFD bag-of-words model was chosen to be 3000 using 'weighted gain ratio'. Other codebook sizes were also investigated to see if better classification accuracy could be achieved. Finally, the best settings for the GFD local approach scored an accuracy of 83.7% in classifying the shapes from the animal shapes dataset, whereas the proposed LPT approach achieved 86% classification accuracy. So we conclude that log-polar transform, as a local descriptor for representing binary shapes, has a clear edge over GFD. Since GFD uses equidistant polar sampling, only rotations in the Cartesian domain are converted to translations in the angular axis (scaling becomes multiplicative). Whereas in the log-polar case, both scale and rotation changes in the Cartesian domain are transformed into translations along the new axes. This invariance to scale and rotation changes gives a clear edge to LPT over GFD, as demonstrated in the experiments above.

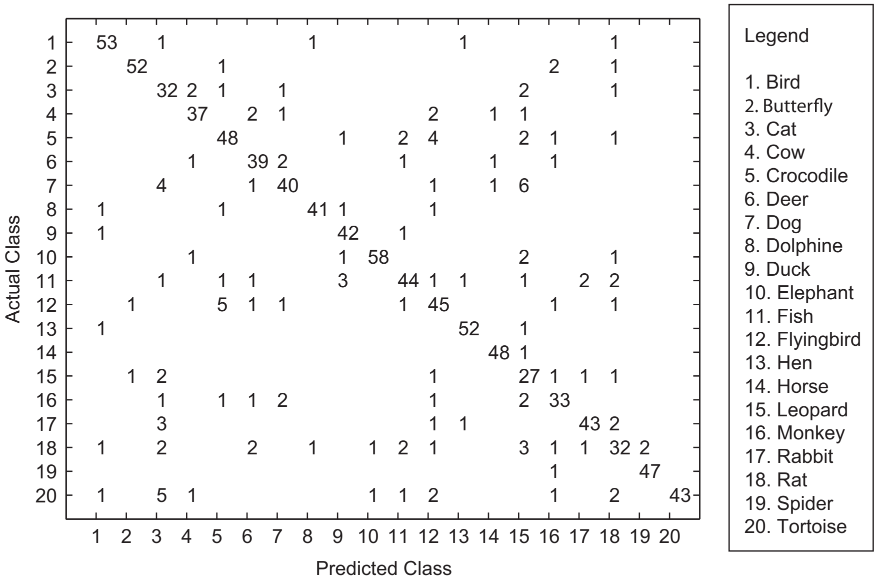
Table 3
Performance comparison of the proposed method with previous works (%).

Method	Accuracy
IDSC [65]	73.6
CS [66]	71.7
CS&SP [52]	78.4
CS&SP&IDSC-F [52]	78.7
CS&SP-DP [62]	80.7
Shape Tree [63]	80.0
HOG-SIFT BoW [27]	80.4
Proposed method	**86.0**

4. Conclusions

We proposed a robust shape classification system, which can handle scale, rotation and strong viewpoint variations, using log-polar transform as a local feature in the bag-of-words framework. In the proposed framework, contextual information was incorporated using a novel method to extract bi-grams from the spatial co-occurrence matrix. We showed that the histogram of bi-gram representation greatly improves on the standard bag-of-words model and its off-shoot, spatial pyramid matching. Besides the above contributions, a novel metric was proposed to select an appropriate codebook size in the bag-of-words model. The selected codebook size was shown to give a high accuracy, compared to a wide range of codebook sizes. Furthermore, the proposed metric for codebook selection is generic, and thus can be used for any clustering quality evaluation. Lastly, we proposed a joint learning framework for learning features in a data-driven manner from the training set. The procedure iterates between setting the codebook size and the maximum radius of the log-polar transform, which was demonstrated to be effective in improving the classification accuracy without the requirement of manual parameter tuning. We tested our algorithm on a challenging shape database and achieved a 6% increase in accuracy compared to state-of-the-art algorithms in the literature.

Our future work would be to extend the framework for object classification in grayscale images. Direct application of log-polar transform as a local feature on every pixel of the image, as in dense SIFT, may not be ideal in terms of computational load or accuracy. An efficient way to perform keypoint detection and feature learning is required, at the least, to be on par with well-established local descriptors like SIFT, LBP, etc.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	53		1					1					1					1		
2		52			1											2		1		
3			32	2	1		1								2			1		
4				37		2	1					2		1	1					
5					48				1		2	4			2	1		1		
6				1		39	2				1			1		1				
7			4			1	40					1		1	6					
8	1				1			41	1			1								
9	1								42		1									
10				1					1	58					2			1		
11			1		1	1			3		44	1	1		1		2	2		
12		1			5	1	1				1	45				1		1		
13	1												52		1					
14														48	1					
15		1	2									1			27	1	1	1		
16			1		1	1	2					1			2	33				
17			3									1	1				43	2		
18	1		2			2		1		1	2	1			3	1	1	32	2	
19																1			47	
20	1		5	1						1	1	2				1		2		43

Actual Class
Predicted Class
Legend
1. Bird
2. Butterfly
3. Cat
4. Cow
5. Crocodile
6. Deer
7. Dog
8. Dolphine
9. Duck
10. Elephant
11. Fish
12. Flyingbird
13. Hen
14. Horse
15. Leopard
16. Monkey
17. Rabbit
18. Rat
19. Spider
20. Tortoise

Fig. 13. Confusion matrix for the best result of our system on the animal shapes dataset.

Table 4
Comparison of Fourier descriptors and global LPT descriptor (%).

	CentroidDistance	GFD	GlobalLPT
Accuracy	35.10	52.80	**53.70**

Conflict of interest

None declared.

Acknowledgments

We would like to thank M.S. Muthukaruppan for engaging in extensive discussion and providing valuable comments. The authors also gratefully acknowledge the anonymous reviewers and editors for their insightful comments which have led to substantial improvements of the paper.

References

[1] F. Mokhtarian, A. Mackworth, Scale-based description and recognition of planar curves and two-dimensional shapes, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (1986) 34–43.
[2] W. Rucklidge, Efficiently locating objects using the Hausdorff distance, Int. J. Comput. Vis. 24 (1997) 251–270.
[3] D. Zhang, G. Lu, A comparative study of fourier descriptors for shape representation and retrieval, in: Proceedings of the 5th Asian Conference on Computer Vision (ACCV), Springer, 2002, pp. 646–651.
[4] H. Freeman, On the encoding of arbitrary geometric configurations, IRE Trans. Electron. Comput. EC-10 (1961) 260–268.
[5] R. Chellappa, R. Bagdazian, Fourier coding of image boundaries, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (1984) 102–105.
[6] J. Livarinen, A. Visa, Shape recognition of irregular objects, in: David P. Casasent (Ed.), Proceedings of the SPIE, Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, vol. 2904, 1996, pp. 25–32.
[7] A. Del Bimbo, P. Pala, Visual image retrieval by elastic matching of user sketches, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 121–132.
[8] H. Asada, M. Brady, The curvature primal sketch, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (1986) 2–14.
[9] G. Dudek, J.K. Tsotsos, Shape representation and recognition from multiscale curvature, Comput. Vis. Image Understand. 68 (1997) 170–189.
[10] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory 8 (1962) 179–187.
[11] A. Goshtasby, Description and discrimination of planar shapes using shape matrices, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-7 (1985) 738–743.
[12] J. Prokop, A.P. Reeves, A survey of moment-based techniques for unoccluded object representation and recognition, in: Graphical Models and Image Processing, CVGIP, vol. 54, Academic Press, Inc., 1992, pp. 438–460.
[13] W.-Y. Kim, Y.-S. Kim, A region-based shape descriptor using zernike moments, Signal Proces.: Image Commun. 16 (2000) 95–102.
[14] D. Zhang, G. Lu, Review of shape representation and description techniques, Pattern Recognit. 37 (2004) 1–19.
[15] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 509–522.
[16] R.A. Messner, H.H. Szu, An image processing architecture for real time generation of scale and rotation invariant patterns, Comput. Vis., Graph., Image Process. 31 (1985) 50–66.
[17] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: In Workshop on Statistical Learning in Computer Vision, ECCV, Springer, 2004, pp. 1–22.
[18] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1265–1278.
[19] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, L. Van Gool, Modeling scenes with local descriptors and latent aspects, in: Tenth IEEE International Conference on ICCV, vol. 1, IEEE Computer Society, 2005, pp. 883–890.
[20] A. Bosch, A. Zisserman, X. Munoz, Scene classification via plsa, in: ECCV, Lecture Notes in Computer Science, vol. 3954, Springer, Berlin, Heidelberg, 2006, pp. 517–530.
[21] S. Agarwal, A. Awan, D. Roth, Learning to detect objects in images via a sparse, part-based representation, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 1475–1490.
[22] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, Int. J. Comput. Vis. 43 (2001) 29–44.
[23] A. Agarwal, B. Triggs, Hyperfeatures—multilevel local coding for visual recognition, in: ECCV, Springer, 2006, pp. 30–43.
[24] K. Grauman, T. Darrell, Efficient image matching with distributions of local invariant features, in: IEEE Computer Society Conference on CVPR, vol. 2, IEEE Computer Society, 2005, pp. 627–634.
[25] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: IEEE Computer Society Conference on CVPR, vol. 2, IEEE Computer Society, 2003, pp. 264–271.
[26] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Tenth IEEE International Conference on ICCV, vol. 2, IEEE Computer Society, 2005, pp. 1800–1807.
[27] K.-L. Lim, H. Galoogahi, Shape classification using local and global features, in: Fourth Pacific-Rim Symposium on Image and Video Technology (PSIVT), IEEE, 2010, pp. 115–120.
[28] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Computer Society Conference on CVPR, vol. 2, IEEE Computer Society, 2006, pp. 2169–2178.
[29] N.M. Elfiky, J. Gonzalez, F.X. Roca, Compact and adaptive spatial pyramids for scene recognition, Image Vis. Comput. 30 (2012) 492–500.
[30] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Computer Society Conference on CVPR, vol. 1, IEEE Computer Society, 2009, pp. 1794–1801.
[31] Y. Jia, C. Huang, T. Darrell, Beyond spatial pyramids: receptive field learning for pooled image features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3370–3377.

[32] Q. Chen, Z. Song, Y. Hua, Z. Huang, S. Yan, Hierarchical matching with side information for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2012, pp. 3426–3433.

[33] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: ECCV, Lecture Notes in Computer Science, vol. 3954, Springer, Berlin, Heidelberg, 2006, pp. 490–503.

[34] J. Li, W. Wu, T. Wang, Y. Zhang, One step beyond histograms: Image representation using Markov stationary features, in: IEEE Conference on CVPR, IEEE Computer Society, 2008, pp. 1–8.

[35] R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Systems, Man Cybern. SMC-3 (1973) 610–621.

[36] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, 1999.

[37] N.A. Bhatti, A. Hanbury, Co-occurrence bag of words for object recognition, in: Proceedings of the 15th Computer Vision Winter Workshop, Czech Pattern Recognition Society, 2010, pp. 21–28.

[38] T. Tuytelaars, C. Lampert, M. Blaschko, W. Buntine, Unsupervised object discovery: a comparison, Int. J. Comput. Vis. 88 (2010) 284–302.

[39] M. Jiu, C. Wolf, C. Garcia, A. Baskurt, Supervised learning and codebook optimization for bag-of-words models, Cognit. Comput. 4 (2012) 409–419.

[40] J. Liu, M. Shah, Learning human actions via information maximization, in: IEEE Conference on CVPR, IEEE Computer Society, 2008, pp. 1–8.

[41] J. Liu, Y. Yang, M. Shah, Learning semantic visual vocabularies using diffusion distance, in: IEEE Conference on CVPR, IEEE Computer Society, 2009, pp. 461–468.

[42] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[43] A.J. Wisniewska, Log-polar Transform, Technical Report, Uniwersytet Zielonogórski, 2004.

[44] D. Young, Straight lines and circles in the log-polar image, in: Proceedings of the 11th British Machine Vision Conference, The British Machine Vision Association (BMVA), 2000, pp. 426–435.

[45] D.G. Lowe, Object recognition from local scale-invariant features, in: IEEE International Conference on CVPR, vol. 2, IEEE Computer Society, 1999, pp. 1150–1157.

[46] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 971–987.

[47] J.-M. Morel, G. Yu, Is sift scale invariant? Inverse Probl. Imaging 5 (2011) 115–136.

[48] I. Kokkinos, A. Yuille, Scale invariance without scale selection, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

[49] A. Andreopoulos, J.K. Tsotsos, 50 years of object recognition: directions forward, Comput. Vis. Image Understand. 117 (2013) 827–891.

[50] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.

[51] C. Xiang, X. Fan, T. Lee, Face recognition using recursive fisher linear discriminant, IEEE Trans. Image Process. 15 (2006) 2097–2105.

[52] X. Bai, W. Liu, Z. Tu, Integrating contour and skeleton for shape classification, in: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2009, pp. 360–367.

[53] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on CVPR, vol. 1, IEEE Computer Society, 2005, pp. 886–893.

[54] A. Vedaldi, B. Fulkerson, Vlfeat: An Open and Portable Library of Computer Vision Algorithms, 2008.

[55] B.G. Mirkin, Mathematical Classification and Clustering, Kluwer Academic Press, 1996.

[56] N. Chinchor, Muc-4 evaluation metrics, in: Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics, 1992, pp. 22–29.

[57] Y. Zeng, J. Tang, J. Garcia-Frias, G. R. Gao, An adaptive meta-clustering approach: combining the information from different clustering results, in: Proceedings of the IEEE Computer Society Conference on Bioinformatics, IEEE Computer Society, 2002, pp. 276–287.

[58] A. Rosenberg, J. Hirschberg, Vmeasure: a conditional entropy-based external cluster evaluation measure, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2007, pp. 410–420.

[59] M. Meila, Comparing clusterings-an information based distance, J. Multivar. Anal. 98 (2007) 873–895.

[60] A. White, W. Liu, Technical note: bias in information-based measures in decision tree induction, Mach. Learn. 15 (1994) 321–329.

[61] J. Feng, B. Ni, D. Xu, S. Yan, Histogram contextualization, IEEE Trans. Image Process. 21 (2012) 778–788.

[62] C. Li, X. You, A. Ben Hamza, W. Zeng, L. Zhou, Distinctive parts for shape classification, in: International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), IEEE, 2011, pp. 97–102.

[63] Y. Li, J. Zhu, F. Li, A hierarchical shape tree for shape classification, in: 25th International Conference of Image and Vision Computing New Zealand (IVCNZ), IEEE, 2010, pp. 1–6.

[64] B. Ni, Learning with contexts (Ph.D. thesis), National University of Singapore, 2010.

[65] H. Ling, D. Jacobs, Shape classification using the inner-distance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 286–299.

[66] K. Sun, B. Super, Classification of contour shapes using class segment sets, in: IEEE International Conference on CVPR, vol. 2, IEEE Computer Society, 2005, pp. 727–733.

[67] G. Granlund, Fourier preprocessing for hand print character recognition, IEEE Trans. Comput. C-21 (1972) 195–201.

[68] D. Zhang, G. Lu, Shape-based image retrieval using generic fourier descriptor, Signal Process.: Image Commun. 17 (2002) 825–848.

[69] R.C. Gonzalez, R.E. Woods, S.L. Eddins, Digital Image Processing Using MATLAB, Pearson Education India, 2004.

B. Ramesh received the B.E. degree in electrical and electronics engineering from Anna University of India in 2009; M.Sc. degree in electrical engineering from National University of Singapore in 2011. Currently, he is pursuing the Ph.D. degree at National University of Singapore under the supervision of Dr. Cheng Xiang and Dr. Lee Tong Heng. His research interests include pattern recognition and computer vision.`

C. Xiang received the B.S. degree in mechanical engineering from Fudan University, China in 1991; M.S. degree in mechanical engineering from the Institute of Mechanics, Chinese Academy of Sciences in 1994; and M.S. and Ph.D. degrees in electrical engineering from Yale University in 1995 and 2000, respectively. He is an Associate Professor in the Department of Electrical and Computer Engineering at the National University of Singapore. His research interests include computational intelligence, adaptive systems and pattern recognition.

T.H. Lee received the B.A. degree (with first-class honors) in engineering from Cambridge University, Cambridge, U.K., in 1980, and the Ph.D. degree from Yale University, New Haven, CT, in 1987. He is a Professor in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests are in the areas of adaptive systems, knowledge-based control, intelligent mechatronics, and computational intelligence.