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:

$$x' = r \cos\theta, \quad y' = r \sin\theta, \qquad (1)$$