Pattern Recognition 67 (2017) 380–395

Contents lists available at ScienceDirect

Pattern Recognition

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

Multiple object cues for high performance vector quantization

B. Ramesh, C. Xiang*, T.H. Lee

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

ARTICLE INFO

Article history:
Received 25 April 2016
Revised 8 December 2016
Accepted 21 February 2017
Available online 22 February 2017

Keywords:
Log-polar transform
Object classification
Visual cues
Bag-of-words model
Flickr-101 dataset
Caltech-101 dataset

ABSTRACT

In this paper, we propose a multi-cue object representation for image classification using the standard bag-of-words model. Ever since the success of the bag-of-words model for image classification, several modifications of it have been proposed in the literature. These variants target to improve key aspects, such as efficient and compact dictionary learning, advanced image encoding techniques, pooling methods, and efficient kernels for the final classification step. In particular, "soft-encoding" methods such as sparse coding, locality constrained linear coding, Fisher vector encoding, have received great attention in the literature, to improve upon the "hard-assignment" obtained by vector quantization. Nevertheless, these methods come at a higher computational cost while little attention has been paid to the extracted local features. In contrast, we propose a novel multi-cue object representation for image classification using the simple vector quantization, and show highly competitive classification performance compared to state-of-the-art methods on popular datasets like Caltech-101 and MICC Flickr-101. Apart from the object representation, we also propose a novel keypoint detection scheme that helps to achieve a classification rate comparable to the popular dense keypoint sampling strategy, at a much lower computational cost.

1. Introduction

Many mammals, especially humans, perceive the world using visual cues as their dominant source of information [1]. Naturally, humans can effortlessly categorize hundreds of objects present in highly complex scenarios using the highly evolved visual cortex that accounts for a variety of visual cues. In fact, many psychophysical studies [2–4] have shown the pre-attentive processing of visual cues, such as color, texture, structure and shape. Thus, we believe a cue-based approach to object categorization is key to achieving real progress toward intelligent systems, and this paper aims to take a step in this direction.

Physiological and clinical studies in humans suggest that visual information processing is highly parallelized, and different cues such as color, depth, form, are perceived by separate channels [5]. Among the visual cues, shape is an elementary aspect of visual processing as it provides important clues about the identity and functional properties of the object. Hence, object recognition research in its budding years was primarily concerned with 3D shape representation, especially in the 80's [6,7]. Nevertheless, due to the representational gap between low-level features and abstract nature of shape components, subsequent two decades of research in object recognition moved away from 3D geometry to appearance-based recognition systems, which opened up new horizons in recognizing natural images [8]. However, in other research areas of computer vision like the saliency detection literature, multiple visual cue processing has been widely adopted (notable examples are [9,10]). Drawing inspiration from the success of the saliency models, we propose parallelized local encoding of multiple object cues like color, structure, texture, and shape, using the log-polar transform in the bag-of-words framework.

The bag-of-words model is a dominant framework in image classification tasks, such as object and scene classification [11–14]. First, keypoint detection [11,15] or dense sampling [16] is done on the image to select patches of interest, followed by a description of each patch using SIFT [11,17], raw patch [18] or filter-based representations [19]. Subsequently, the descriptors are vector quantized using a visual vocabulary or codebook that is commonly built using K-means [11]. Finally, the histograms of the training images are used to train a linear/non-linear classifier.

In the bag-of-words framework, the first big improvement came in the form of spatial pyramid pooling [20], which aims to capture mid-level spatial information by dividing the image into several smaller regions and encoding regional histograms apart from the global bag-of-words histogram representation. Inspired by the spatial pyramid approach, some works have tried to modify its rigid rectangular grid pooling to obtain compact and adaptive representations [21,22]. Besides pooling techniques, much of the effort by the computer vision community has been in the direction of ad-

* Corresponding author.
E-mail address: elexc@nus.edu.sg (C. Xiang).

http://dx.doi.org/10.1016/j.patcog.2017.02.024
0031-3203/

vanced encoding methods, which assign each local descriptor to multiple codewords instead of assigning it to the closest one (vector quantization) or extract covariance measures. Some successful techniques are sparse coding [23], locality constrained linear coding [24], Fisher vector encoding [25], vector of locally aggregated descriptors [26], radial basis coding [27], etc.

The premise of all advanced encoding methods is that information is lost when a local descriptor is simply assigned to the nearest codeword. While improvements have been reported for these advanced encoding methods over vector quantization, high performance improvements have been elusive under controlled conditions [28]. Moreover, the computational cost is rather high for these methods, as noted in [28]. In stark contrast to the previous works, we propose that if the features are powerful enough, vector quantization's performance can be significantly higher than the reported results using various local descriptors such as SIFT, PHOW, self-similarity image descriptor, etc. To this end, we propose a multi-cue log-polar encoded object representation using vector quantization that has significant performance improvement over several encoding methods. While using vector quantization in the bag-of-words model, the classification accuracy relies heavily on the number of local descriptors extracted from the image [29]. Therefore, keypoint selection is an important step.

Recent works using the bag-of-words model for image classification opt for a simple dense sampling strategy instead of a keypoint detector to choose the sampling locations in the image [28,29]. Usually, the number of sampling points chosen by keypoint detectors is drastically outnumbered by the dense keypoints strategy. Consequently, keypoint detectors are often at a disadvantage, as the classification accuracy relies heavily on the number of local descriptors extracted from the image. Although the simplicity of the dense grid keypoints is appealing, the computational cost of extracting and processing the local descriptors is very high. Therefore, we investigate the possibility of a keypoint detection scheme that produces on par performance with the dense keypoint strategy, at a much lower computational cost in terms of both memory and computational time requirements.

In summary, the main contributions of the paper are as follows.

1. A generic object categorization framework is proposed to efficiently combine color, appearance and shape cues using log-polar encoded local descriptors in the bag-of-words model.
2. A novel keypoint detection method is proposed to perform better than the dense sampling strategy from a practical point of view, i.e., the number of local descriptors encoded is much lesser without a significant drop in accuracy. Thus, the proposed keypoint detection scheme using differential entropy offers a more principled approach to image sampling for the popular bag-of-words framework.

Using the proposed features in combination with the simple vector quantization method, the performance of our system is better than several seminal works on the widely tested Caltech-101 dataset and its recently upgraded version, the Flickr-101 dataset. Note that we compare our work to several works that use advanced encoding techniques, or more powerful machine learning paradigms like the multiple kernel fusion, or advanced feature pooling techniques. This implies that if the features are discriminative enough, performance boost can be achieved by simply using vector quantization, which opens up exciting horizons to more powerful machine learning algorithms and encoding methods.

The rest of this paper is organized as follows. We review the related works in Section 2. Then, the details of our proposed methods are introduced in Section 3. Next, the proposed framework is evaluated on two popular datasets and the experimental results are reported in Section 4. Lastly, Section 5 presents the conclusion and future works.

2. Related works

It is commonplace for many works to employ several types of feature descriptors to capture different object cues. For instance, ref. [30] combined dense SIFT, self-similarity descriptors, and geometric blur features with multiple kernel learning to obtain the final image representation. A similar attempt was made in Ref. [31] and Ref. [32] to combine multiple feature channels for image classification. While popular descriptors like SIFT capture texture and gradient information, they do not explicitly encode shape, color and structural object information. However, there are a few hand-crafted local shape descriptors, such as the PHOG shape descriptor proposed in [33], which is a histogram of oriented gradients computed on the output of a Canny edge detector. Likewise, some works [33–35] do obtain local contour fragments to encode shape information from grayscale images, which is a less explored topic of research. In this regard, this paper aims to take a further step to encode object shape, texture, structure and color information in a unified bag-of-features framework using log-polar transform.

One of the primary reasons for adopting log-polar transform in this work is its ability to encode image information without requiring a particular pre-processing step. In other words, it can be used to sample the grayscale image directly or any filtered representation, which simplifies and unifies the task of extracting local descriptors from different object cues. Local shape descriptors employed structurally, such as shape context [36], have been shown to be robust to deformations. It is to be noted that shape context creates log-polar histograms[1] instead of using the classic LPT, which is sampling the image at the intersection of rings and wedges of the transform. Moreover, the shape context requires a point-by-point matching scheme for two shapes, which makes it unsuitable for fast online shape matching [40]. This motivates us to employ the classic log-polar transform (LPT) [41] 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.

One of the first proposals for using log-polar transform as a local feature was by [42]. They developed scale invariant descriptors and used them for object detection. In particular, [42] used bandpass filtered images and transformed the corresponding log-polar sampled amplitude, orientation and phase maps into the Fourier domain. In contrast, we use a log-polar sampling on the structure and texture cues on selected keypoints. In addition, we also combine the idea proposed in [43] to sample the shape boundaries of the extracted binary shape image, by using log-polar transform followed by obtaining its Fourier transform modulus to achieve scale and rotation invariance. This invariant property is another important reason for our choice of log-polar transformation. The selection of keypoints for the structure and texture cues is given below.

In this paper, we define keypoints as visually salient locations in the image. A salient region refers to an area that "stands-out" from its neighborhood and therefore pre-attentively captures attention. In the saliency literature, entropy has been used as a quantifying measure by many works [44–46]. However, entropy based salient detectors like AIM [46] have much lower precision and recall compared to other algorithms developed for salient region detection [10,47] on various benchmark datasets. This means that both the quality and quantity of pixels chosen as salient locations are suboptimal for entropy based saliency detectors. The same observation was confirmed in our experiments using entropy for selecting keypoints in an image, since it does not take into account the shape

[1] Similar trend for grayscale images; popular examples of log-polar histograms are [37–39].