Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Unseen object categorization using multiple visual cues

B. Ramesh*, C. Xiang

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

ARTICLE INFO

Communicated by Liang Lin

Keywords:
Log-polar transform
Object classification
Structure-texture decomposition
Shape extraction
Bag-of-words model
ETH-80 dataset

ABSTRACT

In this paper, we propose an object categorization framework to extract different visual cues and tackle the problem of categorizing previously unseen objects under various viewpoints. Specifically, we decompose the input image into three visual cues: structure, texture and shape cues. Then, local features are extracted using the log-polar transform to achieve scale and rotation invariance. The local descriptors obtained from different visual cues are fused using the bag-of-words representation with some key contributions: (1) a keypoint detection scheme based on variational calculus is proposed for selecting sampling locations; (2) a codebook optimization scheme based on discrete entropy is proposed to choose the optimal codewords and at the same time increase the overall performance. We tested the proposed object classification framework on the ETH-80 dataset using the leave-one-object-out protocol to specifically tackle the problem of categorizing previously unseen objects under various viewpoints. On this popular dataset, the proposed object categorization system obtained a very high improvement in classification performance compared to state-of-the-art methods.

1. Introduction

Object recognition has been a central task to the computer vision community since the early days of using computers to identify hand-written characters [1]. Through these fruitful decades of increasing machine intelligence, we have taken huge strides in solving specific tasks, such as classification systems for automated assembly line inspection [2], hand-written character recognition in mail sorting machines [3], bill counting and inspection in automated teller machines [4], to name a few. Despite these successful applications, computers have made little progress in generalizing object appearance, even under moderately controlled sensing environments. On the other hand, humans can effortlessly categorize hundreds of objects present in highly complex scenarios. This success in pattern recognition is naturally due to the effective utilization of appearance and shape cues, which help in forming distinctive groupings of visual stimuli with different perceivable characteristics [5], by the visual cortex. Therefore, we believe a cue-based approach to object categorization is central to real progress toward intelligent systems and this paper aims to take a step in that direction.

A cue-based approach to object classification is important for generalization to unseen objects. However, this aspect has been rarely studied due to the nature of training and testing protocol used for several object datasets. While the practice of using a random training and testing split avoids the bias of having a fixed training set, it leads to difficulties in objectively assessing whether the training images yield a visual world model that can generalize to unseen objects of a known object category. Another significant obstacle for rigorously evaluating both appearance and shape based methods is the widespread use of databases without segmentation ground truth for the object categorization task.

In this paper, the above problems are addressed by adopting the rigorous leave-one-object-out cross validation protocol on the ETH-80 dataset [6], which provides segmentation ground truth for each object. Even so, unlike previous works on this dataset, we do not make use of the ground truth for classification. In contrast, we propose to extract the shape cue by thresholding [7] the output of a salient object detection model [8]. We only make use of the ground truth for post-mortem analysis of the extracted shape cue, which helps in quantifying the performance of different thresholding methods [7,9–15].

While shape cue provides important clues about the identity and functional properties of the object, several other visual cues assist, both humans and computers alike, in identifying objects from two-dimensional images. Some examples are depth, motion, texture, color, and 3D pose. Nevertheless, object recognition research in its budding years was primarily concerned with 3D shape representation [16,17]. In the late 1980s, the theory of recognition-by-components [18] proposed a powerful set of regularizing constraints using shape primitives for object recognition. It proposed that humans made use of easily detectable perceptual properties (curvature, collinearity, symmetry, parallelism and cotermination) that are invariant to orientation changes, distortion, and occlusion. Nevertheless, this theory has not

* Corresponding author.
E-mail address: bharath.ramesh03@u.nus.edu (B. Ramesh).

http://dx.doi.org/10.1016/j.neucom.2016.12.003
Received 18 July 2016; Received in revised form 18 November 2016; Accepted 2 December 2016

been used successfully in natural images due to the representational gap between low-level features and the abstract nature of model components. Subsequent two decades of research in object recognition moved away from 3D geometry to appearance-based identification systems, which opened up new horizons in recognizing natural images [19].

Although appearance based approaches have taken the forefront of object categorization research [20], contour based object categorization in natural images has been of increasing interest lately [21], with the help of advances in contour detection [22]. This paper aims to take a further step by encoding the object shape and appearance cues in a unified bag-of-features framework using log-polar transform [23]. In particular, we propose a novel fusion of grayscale appearance cues, such as texture and structure, and binary shape information. The input image is decomposed into the structure and texture parts using the Rudin–Osher–Fatemi method [24], and local features are extracted using the log-polar transform on select keypoint locations. As for shape extraction, we employ a state-of-the-art salient object detection model [8] and binarize the saliency map using the classical Otsu's thresholding method [7]. The local shape features are also extracted using log-polar transform at the shape boundaries, following the method proposed in [25]. Finally, each set of the extracted local features (structure, texture, and shape) are fused using the bag-of-words model.

For combining features from different cues, a natural choice for the classification framework is the bag-of-words model, which has become the standard image classification pipeline due to its simplicity and high performance on various datasets [26–28]. The principal idea behind the bag-of-words model is to extract several local features from an image and then identify the likely object from which those features were extracted. The design considerations for each step of the bag-of-words model are discussed below.

Firstly, most works extract local features using a uniform, dense grid of keypoints and report better performance compared to keypoint detection methods [29]. In this work, a novel keypoint detection scheme is used to select the sampling locations and better performance is obtained compared to the dense keypoint strategy. Secondly, heuristically designed local descriptors [30,31] fail to achieve scale and rotation invariant properties theoretically. Therefore, one of the focuses of this paper is to employ a local descriptor with a sound mathematical basis for scale and rotation invariance. In this regard, we employ the log-polar transform [23] for obtaining the local descriptor at each keypoint. Thirdly, the majority of the works quantize the extracted local descriptors using a visual vocabulary or codebook, whose size is chosen in a trial-and-error manner, as noted in [25]. Therefore, we address the issue of choosing the optimal codewords, which aims to reduce the codebook size and simultaneously improve classification performance. In summary, the key contributions of the paper are as follows.

1. An object categorization framework is proposed to efficiently encode appearance and shape cues using the log-polar transform, with very high performance for classifying unseen objects under various viewpoints.
2. A novel keypoint detection method is proposed to select sampling locations using an image denoising method based on variational calculus.
3. An entropy-based codebook optimization scheme is proposed to choose the optimal codewords and simultaneously improve the classification performance.

The rest of the 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 the ETH-80 dataset and the experimental results are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

In this section, we review the related works in each domain of the proposed methods in this paper.

Multi-cue object representation: Several works extract different local descriptors, and treat them as different cues in their object recognition framework. For instance, Khan et al. [32] combined shape cues obtained from SIFT descriptors and color cues obtained from the histogram of sRGB values for object classification. Similarly, Leibe [33] ambitiously combined multiple interest point detectors and multiple descriptors for detecting objects in an image. Likewise, Vedaldi [34] combined dense SIFT, self-similarity descriptors, and geometric blur features with multiple kernel learning to obtain the final image representation. A similar effort was made by [35–37] to combine multiple feature channels for image classification/saliency detection.

Different from the above works, a handful of attempts have been made in the past with the aim of encoding multiple cues by designing a novel image processing method. Ref. [38] combined texture cues obtained from texture-layout filters [39] and contour fragments [40] obtained from sets of edges matched to the image using the oriented chamfer distance. In the same vein, Kumar et al. [41] combined outline contour and the enclosed texture in pictorial structures for object detection. Likewise, some works [21,40,42] obtain local contour fragments to encode shape information from grayscale images. This paper aims to take a further step by encoding the object shape, which is different from encoding the contours of the image, extracted using a salient object detection algorithm [8]. The proposed framework also uses grayscale texture and structure in a unified bag-of-features framework using log-polar transform.

Feature descriptor: Most works adopt the popular SIFT descriptor [31] or shape context [43] for extracting local features. However, other ways of extracting local descriptors, such as filtered responses, normalized pixel values, etc., are also in practice. Normally, dense SIFT descriptors extracted without scale selection are widely reported to give good performance [44], but without the guarantee of the invariant properties. Our work aims to achieve scale and rotation invariance, and in this regard, is most similar to [45], which has used the classic log-polar transform to achieve scale invariance without scale selection for grayscale images. Ref. [45] presents scale invariant descriptors (SIDs) that use a logarithmic sampling on band-pass filtered images. As a result of the non-uniform scale of spatial sampling, centered at each pixel of the image, the authors showed that it is possible to obtain feature vectors that are scale and rotation invariant, by transforming the corresponding log-polar sampled amplitude, orientation and phase maps into the Fourier domain. In contrast to [45], this paper decomposes the grayscale image into the structure and texture filtered images and the resulting cue values are encoded using the log-polar transform at select locations. In addition, we sample the shape boundaries of the extracted binary shape image using the log-polar transform followed by obtaining its Fourier transform modulus.

Keypoint detection: Existing works have adopted two main strategies for selecting keypoints: (1) the simple but counter-intuitive strategy of densely sampling the entire image regardless of object boundaries [44] and (2) the more principled approach of designing sophisticated scale-and-affine invariant keypoint detectors [46]. Our work takes a different approach for selecting keypoints, based on the assumption that a keypoint only needs to be visually salient with respect to its neighbors, and it need not possess invariant properties. Therefore, dealing with noise is a crucial aspect of such a strategy. In this regard, the most related work is in the image denoising literature, which has a multitude of algorithms reviewed extensively in [47]. Gaussian smoothing, anisotropic smoothing (mean curvature motion), total variation minimization, and the neighborhood filters are examples of image denoising methods. Inspired by the success of variational methods on state-of-the-art optical flow benchmark datasets [48,49],