Pattern Recognition Letters 51 (2015) 44–49

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Feature coding for image classification combining global saliency and local difference☆

Shuhan Chen[a,*], Weiren Shi[b], Xiao Lv[c]

[a] College of Information Engineering, Yangzhou University, Yangzhou 225127, China
[b] College of Automation, Chongqing University, Chongqing 400044, China
[c] Chongqing Special Equipment Inspection and Research Institute, Chongqing 401121, China

ARTICLE INFO

Article history:
Received 24 September 2013
Available online 16 September 2014

Keywords:
Image classification
Feature coding
Global saliency
Local difference

ABSTRACT

Saliency[1] based coding proposed recently have been proven to perform well in both performance and efficiency for image classification. However, we find that they are sensitive to unusual features, e.g., noisy features, which we call poor robustness. To address this problem, we propose a novel coding scheme by combining global saliency and local difference together, which are applied for improving stability or robustness and exploring the latent structure information of the codebook respectively. Thorough experiments on various datasets show that our coding consistently performs better than local saliency based coding, in terms of both accuracy and computation cost. Furthermore, it is more robust to unusual features than localized soft-assignment coding. In addition, a combination of our global saliency with local saliency based coding can usually improve both.

1. Introduction

As an important and challenging problem in computer vision, image classification has gained more and more attention in recent years. Many good approaches for image classification have been proposed in the literatures. Among them, the bag-of-words (BOW) [1] model and its extensions (such as spatial pyramid matching [2]) achieve the state-of-the-art performance and have been widely used in many applications. They commonly consist of the following five steps: feature extraction, codebook generation, feature coding and pooling, classification. Feature coding means how to express each descriptor by a codebook to obtain an image-level representation, and has significant influence on classification performance.

We group the existing coding approaches into four categories according to their motivations, as shown in Fig. 1. Voting-based methods are the simplest coding in the literature. Hard-assignment [1] represents a local descriptor to the closest codeword and gives one nonzero coefficient. Without considering codeword ambiguity [3], it always introduces large quantization error. To improve it, soft-assignment [4] is proposed by assigning a local descriptor to all the codewords. Reconstruction-based methods choose a group of codewords to reconstruct descriptors via resolving a least square optimization problem with sparse or locality constraints, e.g., sparse coding [5], local coordinate coding [6], locality-constrained linear coding [7]. Compared with voting-based methods, they always achieve better performance. To reduce reconstruction error, Ren et al. [8] proposed local hypersphere coding, which made reconstruction on a local smooth hypersphere and obtained more distinctive representation. High dimensional methods, proposed for large-scale image classification, such as Fisher kernel coding [9], improved Fisher kernel [10], super vector coding [11], achieve impressive performance [12]. However, they require a large quantity of memory [13]. More recently, saliency based methods are developed, whose core idea is that saliency is a fundamental characteristic of feature coding in the framework with max-pooling [5]. They make a good compromise on efficiency and classification performance. The original salient coding (SaC) [14] encodes each descriptor using the closest codeword by the saliency degree. However, this hard assignment strategy is coarse for feature description [15]. Then, group saliency coding (GSC) [13] is proposed to improve it, whose main idea is calculating the saliency response in a group of codewords. It explores more latent structure information, thus, it performs well.

As mentioned in [15], there are four characteristics we should consider in designing coding method: robustness, adaptiveness, accuracy and independency. Among them, robustness plays the most important role. However, saliency based coding are sensitive to unusual features, e.g., noisy features, in other words, they have poor robustness. In this paper, we propose a novel coding method with good

☆ This paper has been recommended for acceptance by J.K. Kämäräinen.

* Corresponding author. Tel.: +86 18662386487.
E-mail addresses: c.shuhan@gmail.com (S. Chen), wrs@cqu.edu.cn (W. Shi), lvxiao87@126.com (X. Lv).

[1] Saliency in this paper denotes descriptor space saliency.

http://dx.doi.org/10.1016/j.patrec.2014.08.008
0167-8655/

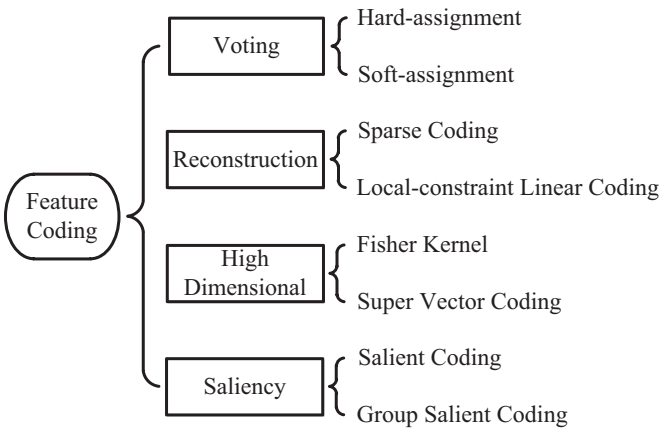


Fig. 1. A taxonomy of coding methods in image classification. Several representatives are listed for each type of coding schemes.

robustness, adaptiveness and independency. Specially, it is achieved by combing global saliency and local difference together, thus, we call it global and local saliency based coding (GLSC). It is noted that they are applied for improving stability or robustness and exploring the latent structure information of the codebook respectively. In addition, our global saliency is complementary to the previous local saliency based coding, thus, a combination can usually improve both.

The remainder of the paper is organized as follows. In Section 2, we briefly review saliency based coding schemes in BOW model. The proposed coding method is presented in Section 3. Section 4 provides experimental results on three datasets: Caltech-101, Scene-15 and UIUC-Sport. Finally, conclusions are drawn in Section 5.

2. Related work

In this section, we mainly concentrate on saliency based coding strategies, introduce their motivations and analyze their limitations. Let x_i ($x_i \in \mathbb{R}^d$) be a d dimensional descriptor, such as scale-invariant feature transform (SIFT) descriptor [16], $B_{d\times M} = (b_1, b_2, \ldots, b_M)$ be a codebook with M cluster centers, and u_i ($u_i \in \mathbb{R}^d$) be the coding coefficient vector of x_i, e.g., u_{ij} be the response of x_i on codeword b_j.

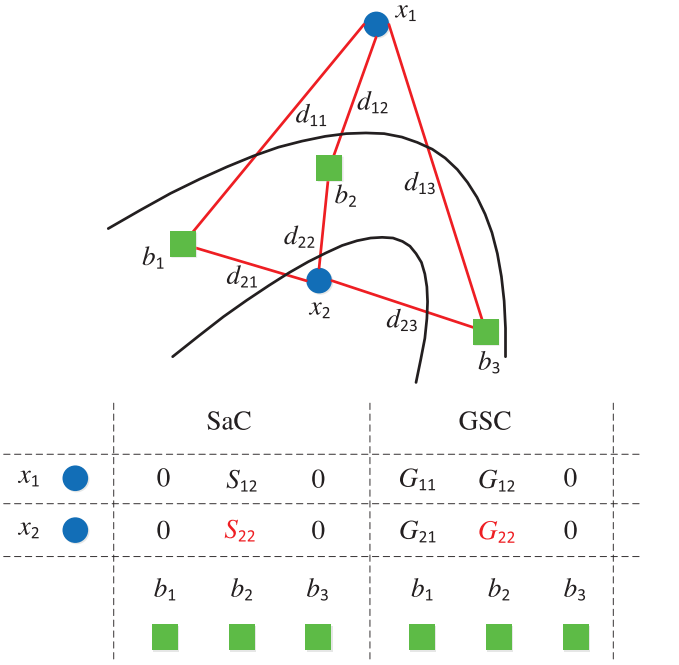
Currently, the framework of using a sparse or local coding scheme combining with max-pooling is regarded as the state-of-the-art. Pooling operation is used to obtain an image-level representation. In the max-pooling, only the strongest response will be preserved. Let p_j be the ith component of image representation p, then max-pooling can be defined as:

$$p_j = \max_i u_{ij} \tag{1}$$

where $i = 1, 2, \ldots, l$, and l is the total number of local features in an image. A detailed analysis of feature pooling was conducted in [22], including average [1], sum [2], max pooling, we only concentrate on max-pooling in this paper.

In saliency based coding, a strong response on a codeword means that this codeword is much closer to a descriptor belonging to it comparing with the other codewords [15]. It indicates the codeword can represent this descriptor independently, which is measured by saliency degree in saliency based coding. In the original salient coding, it is defined by measuring the difference between the closest code and other K-1 codes. In detail, a descriptor is represented as:

$$u_{ij} = \begin{cases} \Psi(x, b_j), & \text{if } j = \arg\min_j \left(\|x - b_j\|^2\right) \\ 0, & \text{otherwise} \end{cases}$$
$$\Psi(x_i, \tilde{b}_j) = 1 - \frac{\|x_i - \tilde{b}_j\|_2}{[1/(K-1)]\sum_{k\neq j}^{K} \|x_i - \tilde{b}_k\|_2} \tag{2}$$



$$S_{12} = 1 - \frac{d_{12}}{d_{11}} > S_{22} = 1 - \frac{d_{22}}{d_{21}}$$
$$G_{12} = (d_{11} - d_{12}) + (d_{13} - d_{12}) > G_{22} = (d_{23} - d_{22}) + (d_{21} - d_{22})$$

Fig. 2. Illustration of saliency based coding. The blue balls are local descriptors and the green rectangles are codewords. The red lines denote the Euclidean distance between them in descriptor space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where Ψ denotes the saliency degree, and is the set of K closest codewords to descriptor x.

Although performs well in both effectiveness and efficiency, there still exists a limitation caused by the coarse hard assignment strategy. Only considering the closest codeword, the representations of some descriptors may be suppressed in the subsequent max-pooling. Take Fig. 2 for example, wherein x_1, x_2, x_3 and b_1, b_2, b_3 denote local descriptors and codewords respectively, S_{ij} and G_{ij} denote the response of x_i to b_j in SaC and GSC respectively. As described in Fig. 2, S_{22} is suppressed by S_{12} (since $S_{22} < S_{12}$), thus, it will lose the representation of descriptor x_2.

To improve it, Wu et al. [13] proposed GSC method by introducing group coding. Its main idea is to compute the saliency response in a group of codewords with different group code sizes, and the maximum of all responses is preserved in the final coding result. Let v^g denote the coding response with group code size g, then the GSC representation can be described as:

$$u_{ij} = \max\{v_{ij}^g\}, g = 1, \ldots, G$$
$$v_{i,j}^g = \begin{cases} \Phi^g(x_i), & \text{if } b_j \in g(x_i, g) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$
$$\Phi^g(x_i) = \sum_{t=1}^{G+1-g} \left(\|x_i - \tilde{b}_{g+t}\|_2 - \|x_i - \tilde{b}_g\|_2\right)$$

where in Φ^g denotes group saliency degree, $g(x_i, g)$ denotes the set of the g closest codewords of descriptor x_i, and G is the maximum group code size.

GSC not only preserves the good properties of effectiveness and efficiency in SaC with the help of group coding, but also performs more stably and robustly than SaC. Consider the example in Fig. 2 ($G = 2$), although the response G_{22} is also suppressed by G_{12} (since $G_{22} < G_{12}$), we can still find the representation of descriptor x_2 on