Pattern Recognition Letters 46 (2014) 83–88

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Exploring neighborhood and spatial information for improving scene classification ☆

Kelly Assis de Souza Gazolli *, Evandro Ottoni Teatini Salles

Electrical Engineering Department of Universidade Federal do Espírito Santo, Av. Fernando Ferrari, Goiabeiras, 29060-900 Vitória, ES, Brazil

ARTICLE INFO

Article history:
Received 14 October 2013
Available online 2 June 2014

Keywords:
Image descriptor
Contextual information
Non-parametric transform
Spatial arrangement information
Scene classification

ABSTRACT

A good image descriptor is essential for the scene classification task. This work proposes an improvement for the Contextual Mean Census Transform (CMCT), an image descriptor, obtained by adding information from distant neighbors to the non-parametric transform computation and spatial information. We also combine the new descriptor with the gist and Spatial Modified Census Transform (Spatial MCT) descriptors to improve classification performance. Experimental results on four commonly used datasets demonstrate that the proposed methods can achieve performance competitive with previous methods.

1. Introduction

Scene classification is a very popular topic in the field of computer vision finding many applications, such as, content-based image organization and retrieval [31,4], automatic creation of photo albums and robot navigation [28]. Although humans can understand a real-world scene quickly and accurately, automatic scene classification is not an easy task because of, among other reasons, the high intraclass variety, the ambiguities in some scenes, as well as the variations of illumination and scale.

An important step in scene classification is the image representation. Traditional scene classification approaches use low-level global features [11,27], such as color and texture distributions of the pixels. However, using such features, these methods have been employed to classify amongst a small number of image categories, usually, 2.

Oliva and Torralba [21,22] showed that scenes which belong to the same category, normally, have the same spatial layout properties and proposed a holistic approach for building the gist of the scene from global features. The low dimensional global features are based on configurations of spatial scales and estimated without invoking segmentation or grouping operations [14]. Nevertheless, if scenes with similar global characteristics are to be differentiated, then global features may not be discriminative enough [24].

A popular approach for scene representation is the bag-of-words [5,30,2], inspired by the bag of words model used in text categorization. In such an approach, a set of local image patches is, generally, dense sampled. Next, a visual descriptor is extracted, normally using the SIFT descriptor [17], and quantized into one visual word. Application of this method results in representation of each image by a histogram of visual words. Many variants of this model have been proposed. Lazebnik et al. [13] proposed a spatial pyramid, a technique which works by partitioning the image into increasingly fine sub-regions. Quelhas et al. [26] showed that a text-like bag-of-visterms (histogram of quantized local visual features) is suitable for scene classification and used probabilistic latent semantic analysis (pLSA) to find intermediate topics. In [7], latent variables are learned using Latent Dirichlet Allocation. Qin and Yung [24] proposed a method based on contextual visual words, in which the contextual information from the coarser scale and neighborhood regions to the local region of interest are included. Li et al. [16] also proposed a contextual bag-of-word model similar to the approach proposed in [24], but with different implementation.

Although capable of good results, the bag-of-words approach has some disadvantages. For example, the codebook should be large enough so that each image can be properly represented by a histogram [33], which makes the codebook size dependent on the dataset. Furthermore, the codebook-building process is often computationally intensive, which limits its efficiency [33].

Taking another direction, some works are adopting the Census Transform [37], a non-parametric transform, also known as Local

☆ This paper has been recommended for acceptance by Eckart Michaelsen.

* Corresponding author. Permanent address: Instituto Federal do Espírito Santo, Rodovia ES-010, Km 6.5, Manguinhos, 29173-087 Serra, ES, Brazil. Tel.: +55 27 4009 2169; fax: +55 27 4009 2644.

E-mail addresses: kasouza@ifes.edu.br (K. Assis de Souza Gazolli), evandro@ele.ufes.br (E. Ottoni Teatini Salles).

http://dx.doi.org/10.1016/j.patrec.2014.05.009
0167-8655/

Binary Pattern (LBP) [20], in the image representation. Wu and Rehg [35] proposed CENTRIST (Census Transform Histogram), a holistic representation that captures structural properties, rough geometry and generality by modeling distribution of local structures through a Census Transform histogram. This method has the following advantages: easy implementation, no parametric, very low computational cost and invariance to illumination. Song and Li [29] proposed using hierarchical features for image representation by exploiting the combined strengths of the wavelet transform and LBP. Gazolli and Salles [10] proposed CMCT (Contextual Mean Census Transform) which combines the distribution of local structures with contextual information.

In this paper, we propose an extension of CMCT, ExtendedCMCT (ECMCT), by adding information from the neighbors in a window of pixels, when computing the non-parametric transform, that are close, but which are not boundary neighbors of this window. For avoiding a dimension explosion, we only use some selected pixels placed at a distance of k pixels from the window center. We also include information on the mean of contrast and mean and variance of the non-parametric transform values, computed on sub-regions in the image, with the intent of providing some clues of the spatial arrangement of features in the image.

The proposed descriptor is combined with other techniques for improving the scene classification performance. Experiments have shown that the proposed methods effectively improve the results of scene representation.

The rest of this paper is organized as follows: Section 2 presents some aspects of CMCT. Section 3 details the proposed approaches. In Section 4, we report the classification performance of our methods on four different datasets comparing it with some approaches previously reported in the literature and, finally, in Section 5, the conclusion is presented.

2. CMCT – Contextual Mean Census Transform

Contextual Mean Census Transform (CMCT) [10] is an image descriptor inspired by CENTRIST [35] which adds contextual information to local structures. The main idea is differentiating windows that have similar structures, but are inserted in neighborhoods significantly different. For accomplishing this task, CMCT creates a new local structure from the local structure of a 3×3 window of pixels and from the local structures of its neighboring windows.

The CMCT computation adopts MCT (Modified Census Transform) [9], a nonparametric transform, which is obtained in the following manner: first, the Modified Census Transform, $MCT(x,y)$, computes a mean $\bar{I}(x,y)$ over 3×3 window of pixels. So, every pixel in the 3×3 window is then compared with $\bar{I}(x,y)$. If the pixel is larger than or equal to $\bar{I}(x,y)$, a bit 1 is set in the corresponding location, otherwise, a bit 0 is set, as follows:

$$MCT(x,y) = \otimes_{(i,j)\in\mathcal{N}'(x,y)}\zeta(I(i,j),\bar{I}(x,y)),$$
$$\zeta(m,n) = \begin{cases} 1, & m \geqslant n \\ 0, & m < n \end{cases} \quad (1)$$

where $\otimes$ represents concatenation operation, $\bar{I}(x,y)$ is the mean of the intensity values in the 3×3 window of pixels centered at (x,y), $I(i,j)$ is the gray value of the pixel at (i,j) position, $\mathcal{N}'(x,y) = \mathcal{N}(x,y) \cup (x,y)$ and $\mathcal{N}(x,y)$ defines a local spatial neighborhood of the pixel at (x,y), so that $(x,y) \notin \mathcal{N}$. In the Modified Census Transform technique, 9 bits are generated and converted to a decimal number in [0, 511], the MCT value.

The MCT used in CMCT differs slightly from the original, because $\bar{I}(x,y)$ is not compared with the center pixel. Thus, MCT generates 8 bits, instead of 9, which are converted to a decimal number in [0, 255]. In order to differentiate Modified Census Transform with 9 bits from Modified Census Transform with 8 bits, the latter is referred as MCT8.

The steps for CMCT generation are as follow. First, MCT8 is computed for all pixels in the image. Then, a histogram of MCT8 values is obtained. A new image is created in which the original image pixels are replaced by the correspondent MCT8 values. In the sequel, MCT8 is computed on the new image pixels and a new histogram is generated. Then, the MCT8 histogram for the original image and the MCT8 histogram for the new image are concatenated, generating the CMCT descriptor.

3. The proposed approach

The proposed descriptor, ExtendedCMCT, is an extension of the CMCT descriptor and improves the representation effectiveness of CMCT by adding two different information types: neighborhood information and spatial arrangement information.

3.1. Extracting information from distant neighbors

The addition of neighbors information improves representation capacity of MCT8 [10]. To increase the image representation efficiency of CMCT, we include information about neighbors that are not considered in the original proposal. The information from distant neighbors increments the contextual information of a 3×3 window of pixels and helps even more in differentiating windows that are similar, but are placed in different regions, increasing, in this way, the region size from which the local information is extracted. The idea is to create a new structure composed by the pixels of the current window and pixels positioned in regions that are close, but not boundary neighbors of this window.

A straightforward manner for including this kind of information is to increase the window size for the computation of MCT8. However, this approach also increases, in a remarkable way, the final feature vector size. If a window of 5×5 is considered, for example, the size of the MCT8 values histogram will be 2^{24} positions. In consideration of this, intending to avoid the excessive increase of the descriptor size, only eight pixels positioned at k pixels distant from the central pixel of the current window are considered, as illustrated in Fig. 1. The mean over the 3×3 window, $\bar{I}_w(x,y)$, is computed and, then, the mean of the 8 pixels and $\bar{I}_w(x,y)$, $\bar{I}(x,y)$, is calculated. The 8 pixels are compared to $\bar{I}(x,y)$. The Census Transform of Distant Neighbors (CTDN) value at a distance k from the pixel at position (x,y), $CTDN_k(x,y)$, is calculated as follows:

$$CTDN_k(x,y) = \otimes_{p=0}^{p=7}\zeta(N_p,\bar{I}(x,y)),$$
$$\zeta(m,n) = \begin{cases} 1, & m \geqslant n \\ 0, & m < n \end{cases} \quad (2)$$

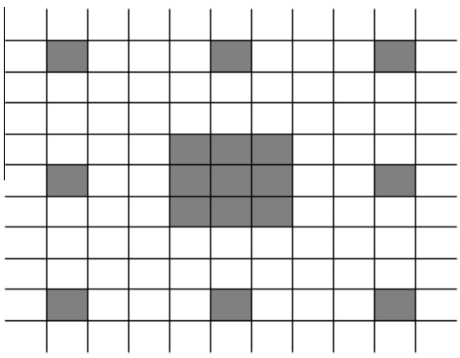

Fig. 1. The pixels considered in the CTDN calculation for $k = 4$.