Computer Vision and Image Understanding 116 (2012) 54–67

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Low-dimensional and comprehensive color texture description

Susana Alvarez [a,*], Anna Salvatella [c], Maria Vanrell [b,c], Xavier Otazu [b,c]

[a] Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Campus Sescelades, Avinguda dels Països Catalans, 26, 43007 Tarragona, Spain
[b] Dept. Ciències de la Computació, Universitat Autònoma de Barcelona, Spain
[c] Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain

ARTICLE INFO

Article history:
Received 19 November 2009
Accepted 29 August 2011
Available online 2 September 2011

Keywords:
Color texture descriptors
Basic terms vocabulary
Retrieval
Segmentation
Browsing

ABSTRACT

Image retrieval can be dealt by combining standard descriptors, such as those of MPEG-7, which are defined independently for each visual cue (e.g. SCD or CLD for Color, HTD for texture or EHD for edges). A common problem is to combine similarities coming from descriptors representing different concepts in different spaces. In this paper we propose a color texture description that bypasses this problem from its inherent definition. It is based on a low dimensional space with 6 perceptual axes. Texture is described in a 3D space derived from a direct implementation of the original Julesz's Texton theory and color is described in a 3D perceptual space. This early fusion through the blob concept in these two bounded spaces avoids the problem and allows us to derive a sparse color-texture descriptor that achieves similar performance compared to MPEG-7 in image retrieval. Moreover, our descriptor presents comprehensive qualities since it can also be applied either in segmentation or browsing: (a) a dense image representation is defined from the descriptor showing a reasonable performance in locating texture patterns included in complex images; and (b) a vocabulary of basic terms is derived to build an intermediate level descriptor in natural language improving browsing by bridging semantic gap.

1. Introduction

Due to the growth in size of image collections and the need to retrieve semantically-relevant images from them, the development of effective systems for image retrieval has acquired great importance since the early 1990s. Since then, the studies on the development of content-based image retrieval systems have widely increased. The goal of these content-based image retrieval (CBIR) systems is to represent and to index image databases using the visual content of the images such as color, shape, texture and spatial layout, so low-level image feature extraction is the basis of CBIR systems. Usually multi-dimensional feature vectors are used to describe these contents. The descriptors can either be extracted from the entire image or from regions. In the first case, the image is often characterized by its histogram thus obtaining a global image description. In the second case, image regions are obtained partitioning the image into tiles from which features are extracted; this is a way of representing the global features of the image at a finer resolution [1,2]. The most important drawback to extract image visual content of both methods has been the inability to capture semantic content.

A better method to obtain regions is to use segmentation algorithms to divide images into homogeneous regions according to some criteria that discriminate between different entities of the image. This is the first step of all region-based image retrieval systems (RBIR). Then some descriptors are defined so that the retrieval can be performed [3–7]. These methods have significantly improved retrieval results, but they are still different from the results obtained by humans.

The main problem of current retrieval systems simulating the search performed by a human subject is the difference between human description of the queried image and the level of description (the extracted information) of retrieval system. Human subjects use high level concepts (and words) to identify elements of the image, actions or situations, whereas retrieval methods extract low level features (i.e. color, texture, shape, etc). The difference between these description levels is known as the 'semantic gap' [2]. One way to reduce the 'semantic gap', pointed out by Liu et al. [8] in their survey on CBIR systems, is the use of object ontology to define high-level concepts. This requires to obtain objects/entities of images. Some works have studied this issue in narrow application domains [9–12]. Another way would be to define descriptors presenting the image components in linguistic terms, which is one of the goals of this paper.

Recently, the bag-of-words model uses image features as 'visual words' [13] of a wide vocabulary, mapped onto image categories by machine learning techniques [14]. The learning process deals with the whole width of the semantic gap. These approaches achieves important results in general categorization of scenes or objects

* Corresponding author. Fax: +34 977 559610.
E-mail address: susana.alvarez@urv.es (S. Alvarez).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.08.004

even when the vocabulary is based on low-level features. One question that arise from our work is how these techniques could improve the results by introducing more semantic information in their vocabularies.

In the specific cases of color and texture, the most usual descriptors are low-level features combined with shape or spatial location features. Descriptors are sometimes obtained from histograms [3,15,9,16]. Other color descriptors capture the spatial color distributions: color layout (CLD) and color structure (CSD) descriptors. These last descriptors and descriptors obtained from histograms are included in the MPEG-7 [17] as standard color descriptors. In regard to texture descriptors, there are different sets of features, for example, wavelet features using Gabor filters [15,18,17,7] or rotated complex wavelet filters [19], both define the multiscale descriptor as a vector containing energy and energy deviations before the corresponding filter is applied to the image. Liu and Picard [4] developed the 'Wold' features which distinguish between 'structured' and 'random' texture components. The former correspond to the peak magnitudes of image autocovariance and the latter are the MRSAR (Multiresolution simultaneous autoregressive model) estimated coefficients. Barcelos et al. [20] define a texture descriptor based on the modal matrix that represent the frequency space of an image consisting of eigenvectors that measure the proximity among points set of the quantized power spectrum of image. The modal matrix is their texture descriptor. Zhong and Jain [21]'s color and texture descriptor is a vector that contains some coefficients of the DCT (Discrete Cosine Transform) in JPEG image format. Lazebnik et al. [22] defined the *RIFT* descriptor as an sparse representation of the *SIFT* [23] that tries to cope with image textons assuring rotation invariance. All of these descriptors do not directly map the set of properties they extract to words describing the image.

If we focus on the problem of descriptors that can be mapped to real words, few descriptors have been developed. Most of them are generally related to color properties. Carson et al. [24] extracts two dominant colors from each region; Mojsilovic et al. [25] and Ma and Manjunath [5] from different codebooks, build feature vectors with the dominant colors and its corresponding occurrence percentage within the image. Smith and Chang [26], using a sparse binary vector representation of color sets, allow users to specify the color content within images by picking colors from a color chooser or by textual specification. Finally, Benavente et al. [27] proposed a fuzzy set model that directly maps colors to the eleven English basic color names. In the case of texture descriptors mapping words, Manjunath et al. [28] developed the PBC, which consist of three perceptual features: regularity, directionality and scale represented by bounded values. These features are related to the three most important perceptual dimensions in natural texture discrimination 'repetitiveness, directionality and granularity' identified by Rao and Lohse [29] in a psychophysical experiment. Recently, Salvatella and Vanrell [30] proposed a sparse texture descriptor that is based on describing texture through their blob attributes, this is the starting point of the proposal of this paper.

Focusing on the previous idea of mapping descriptors onto words we founded more recent works on image annotation [31–34], these works follow a top-down methodology essentially based on machine learning techniques. The main focus relies on the accuracy on predicting good annotations by learning from previously annotated images, usually based on standard descriptors commonly used.

Here in this work we go back to the descriptor definition step by proposing a compact descriptor called *Texture Component Descriptor*, which deals with the annotation of color-textures without any learning step. Our descriptor relies on a pure bottom-up approach where feature selection is inspired on perceptual assumptions. We justify this backtracking to the descriptor definition because we can achieve two desired properties: the descriptor is low dimensional and comprehensive. This is, it is based on six dimensions with a direct perceptual correlation each. These properties can be achieved since we substitute machine learning effectiveness by strong perceptual assumptions. These are directly derived from the texton theory [35] which is complemented with perceptual grouping mechanisms capturing patterns emerging from the repetition of local attributes [36].

The paper is organized as follows: in Section 2 we review the perceptual considerations justifying the attribute space where the descriptor is based on. In Section 3 we propose a descriptor *Texture Component Descriptor* (TCD) derived from a 6D space that is an early fusion of a 3D blob space and a 3D color space. The next sections will explore the comprehensive nature of the proposed descriptor: in Section 3.1 we propose a dense image representation for image segmentation, and in Section 4 we define a grammar that translate our descriptor to basic linguistic terms that can improve it in browsing applications. Afterwards, Section 5 compiles all the experiments that evaluates our approach. The first experiment demonstrates that our descriptor achieves similar performance to current best descriptors in retrieval; we compare our TCD to MPEG-7 in standard Corel datasets. Subsequent experiments explore the behavior of the descriptor from a qualitative point of view showing its feasibility in segmentation and browsing applications. In the last section we summarize the proposal and outline further work.

2. Texture and blobs

Texture representation has been the focus of a large amount of research in Psychophysics [37,36,35,29] too. Two different schools of thought in the study of texture segregation have converged in their final conclusions. Both first-order statistics of local features and global spatial considerations are needed for a full representation. The present work is based on the texton theory of Julesz and Bergen [35] as the basis for the first steps in texture perception. After different conjectures, in 1983 Julesz proposed this theory that states three heuristics. First, texture discrimination is a preattentive visual task. Second, textons are the attributes of elongated blobs, terminators and crossings. Third, preattentive vision directs attentive vision to the location where differences in density of textons occur, ignoring positional relationships between textons. Finally, he gives an explicit example of textons in this way: *"elongated blobs of different widths or lengths are different textons"*. In summary, Texton theory concludes that preattentive texture discrimination is achieved by differences in first-order statistics of textons, which are defined as line-segments, blobs, crossings or terminators and their attributes: width, length, orientation and color.

This perceptual theory is the consequence of an exhaustive study on local texture properties provoking preattentive texture discrimination in experimental conditions. In this work we propose to use these powerful results derived from a large psychophysical experimentation trying to prove different conjectures. These results allow us to substitute the usual training step on annotated image datasets of most computational approaches. Our hypothesis is based on the fact that these perceptual features can be encoding the efficiency of human visual representation. With the same goal, an early computational implementation of texton theory was done by Voorhees and Poggio [38], blob attributes on gray level images were used to determine boundaries between textures. In this work we propose to continue the work of Voorhees and Poggio [38] by updating it with recent computational operators [39] using color attributes [40] and inserting one further step that simulate a grouping mechanism onto the attributes that captures emergent repetitiveness [36].

Apart from the assumption that a texture can be described by their blob attributes, we also assume that a texture is provided