Perceptual color texture codebooks for retrieving in highly diverse texture datasets

Susana Alvarez
Dept. of Computer Science & Mathematics
Universitat Rovira i Virgili
Tarragona, Spain
susana.alvarez@urv.es

Anna Salvatella, Maria Vanrell, Xavier Otazu
Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain
maria.vanrell@cvc.uab.es

***Abstract*—Color and texture are visual cues of different nature, their integration in a useful visual descriptor is not an obvious step. One way to combine both features is to compute texture descriptors independently on each color channel. A second way is integrate the features at a descriptor level, in this case arises the problem of normalizing both cues.**

A significant progress in the last years in object recognition has provided the bag-of-words framework that again deals with the problem of feature combination through the definition of vocabularies of visual words. Inspired in this framework, here we present perceptual textons that will allow to fuse color and texture at the level of *p-blobs*, which is our feature detection step. Feature representation is based on two uniform spaces representing the attributes of the *p-blobs*. The low-dimensionality of these texton spaces will allow to bypass the usual problems of previous approaches. Firstly, no need for normalization between cues; and secondly, vocabularies are directly obtained from the perceptual properties of texton spaces without any learning step. Our proposal improve current state-of-art of color-texture descriptors in an image retrieval experiment over a highly diverse texture dataset from Corel.

***Keywords*-color-texture; descriptor; perceptual; retrieval;**

I. Introduction

In the literature we can find several works dealing with color and texture in different applications, but the integration of both cues is still an open issue. In this regard, there have been several ways of integrating these features. In some of them, [1], [2] color and texture are processed separately, and then, they are combined at the similarity measure level. This means that for every visual cue a dissimilarity measure is obtained, each one in a different space, and then they are combined to obtain a final similarity that needs to be scaled in order to be comparable. Other authors [3], [4], [5] use the same descriptor over each component of a color space, cue combination is done jointly at the level of the color channel whose outputs are concatenated afterwards.

In this paper we propose a color-texture descriptor that does not directly follow any of these two previous approaches. Color and texture cues are combined through the blob concept as a direct implementation of the original definition of texton given by Julesz [6]. Texton theory gives the basis for the first steps in texture perception. Julesz argued that texture discrimination is achieved due to differences in density of textons. Textons are the attributes of elongated blobs, terminators and crossings. 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. We will not consider terminators or crossings, since it is not clear it would be necessary for natural images. The descriptor we propose is a first-order statistic of the attributes of the blobs clearly perceived in the images or *perceptual blobs (p-blobs)*.

This texton definition for texture representation, that we are computationally approaching in this work, matches perfectly with current bag-of-word models [7] arisen from the object recognition field. In these models image representation is built in three main steps: feature detection, feature description and vocabulary generation; no matter where the features are located but their frequency. In our proposal, feature extraction is done through the detection of *p-blobs*. Feature representation or description is our main contribution, we compute the shape and color attributes of *p-blobs* in two uniform spaces. These spaces are low-dimensional and the perceptual properties of their axes will allow a direct generation of the color and shape vocabularies which are lately fused to form the final descriptor.

Our proposal presents two main advantages: firstly, we work with spaces which are bounded, low-dimensional and perceptual, in this way there is no need of specific normalization between cues. Secondly, the perceptual properties of the spaces allow a direct and meaningful generation of vocabularies without any learning step and in a feedforward manner [8]. The direct generation of the vocabularies bypasses the problem of codebook construction that is still an open problem to fuse color and spatial features [9], [10]. In the last sections we show how our descriptor improves current state-of-art of color-texture descriptors in an image

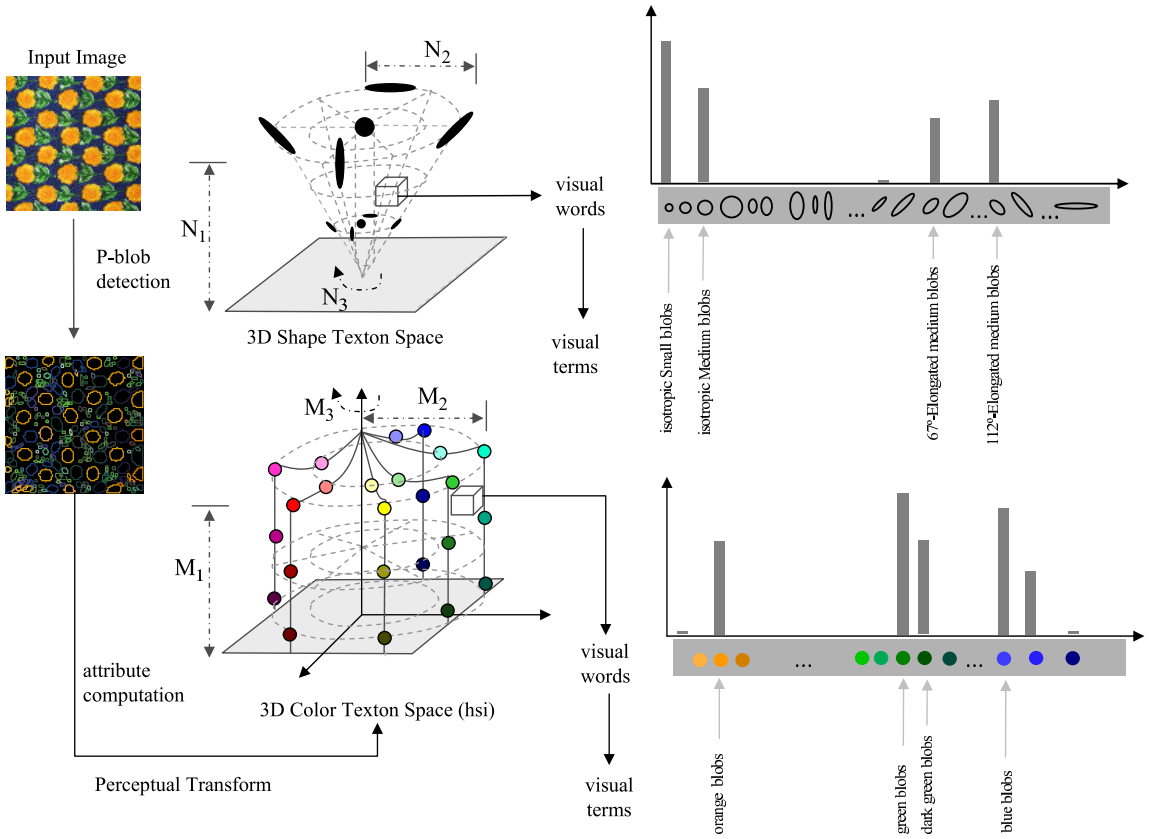


Figure 1. Perceptual Texton Descriptor (PTD)

retrieval experiment over a highly diverse texture dataset from Corel.

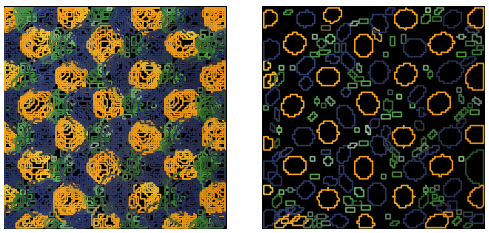
II. Perceptual Texton Codebook

In this section we introduce the steps to build our descriptor, illustrated in figure 1. To this end we will first introduce the *p-blob* concept, afterwards we show how to compute their attributes and finally we explain how to build the vocabulary that derives the final descriptor.

A. Perceptual blobs

For an early detection of blobs of an image we use the approach in [11], which is based on the differential operators in the scale-space representation. Blobs are detected with the normalized differential Laplacian of the Gaussian operator and refined with a windowed second moment matrix. Since blob information emerge from both intensity and chromaticity variations, we apply this procedure to each component of the opponent color space. To be invariant to intensity changes, previously the components have been normalized.

This first step generates a large amount of redundant blobs, we can see the example in Fig.2.(a), that does not match the aim of the perceived blobs. To obtain the *p-blobs* we apply a filtering stage where a winner-take-all competition among overlapping blobs is done. *P-blobs* obtained in the previous example are shown in figure 2.(b).



(a) (b)

Figure 2. (a) Detected blobs. (b) p-blobs

B. Attribute Computation

In the previous step we have obtained a list of blobs, whose shape attributes are computed in the detection step, these are: width, length and orientation of the *p-blobs*, denoted as (w, l, θ). The shape attributes of the whole image are consequently given in matrix form by $\mathbf{T_{sha}} = [\mathbf{WL\Theta}]$.

These shape attributes are transformed to a new space, we call the *shape texton space*, that has been built taking into account perceptual considerations on the appearance of image blobs. This perceptual shape texton space is obtained with the following non linear transformation U,

$$\begin{array}{rccc} U: & \mathbb{R}^3 & \rightarrow & \mathbb{R}^3 \\ & (w,l,\theta) & \rightarrow & (r,z,\phi) \end{array} \quad (1)$$

where $r = \log_2(ar), z = \log_2(log_2(A))$ and $\phi = 2\theta$, being

ar the blob aspect ratio ($ar = w/l$), A its area ($area = w \cdot l$) and θ its orientation.
Shape attributes are then represented in cylindrical coordinates, in the same way as color attributes of the *p-blobs*. These are represented in the HSI color space because it is perceptual and has similar properties to uniform color spaces. Hence, color attributes of the *p-blobs* will be computed as the median values through the *p-blob* area of the three coordinates, (h, s, i), that are hue, saturation and intensity, respectively. For the whole image, it is given in matrix form by $\mathbf{T_{col}} = [\mathbf{H\,S\,I}]$.

C. Vocabulary Construction

To build the texton vocabulary we will directly sample the perceptual spaces. Visual words will be related to a combination of blob attributes. These sampling is done specifically for each texton space. For the case of the shape texton space we will build $N_1 \times N_2 \times N_3$ visual words, which correspond to the number of bins on the area, the aspect-ratio, and the orientation axis, respectively. For the case of the color texton space we build $M_1 \times M_2 \times M_3$ visual words corresponding to luminance, saturation and hue, respectively.

The proposed vocabulary brings to a unique correspondence between a visual word and a specific set of attributes for the *p-blob* appearance, avoiding redundancy and having visual words with perceptual meaning. An example of the basic visual terms that can be associated to the visual words of the vocabulary are shown in figure 1 for a given image. In this figure we show the overall scheme to build the perceptual texton bag-of-words descriptor, and its interpretation.

The experiments shown in next section have used a total amount of 424 visual words, from the addition of 168 visual words from the shape sampling $(N_1, N_2, N_3) = (7, 3, 8)$, and 256 visual words from the color sampling $(M_1, M_2, M_3) = (4, 4, 16)$. The Perceptual Texton Descriptor (*PTD*) is formed by the concatenation of the two texton occurrences. As in previous works [12], [13], we measure similarity between these histograms using the standard χ^2-*distance* distribution distance.

III. Experiment

We evaluate the performance of the PTD descriptor in an image retrieval experiment. We have used six different datasets from the Corel stock photography collection[1]: Textures (137000), Textures II (404000), Various Textures I (593000), Various Textures II (594000), Textile Patterns (192000) and Sand & Pebble Textures (390000). In the experiment we refer to them as *Corel*, *Corel2*, *CorelV1*, *CorelV2*, *CorelTex* and *CorelSan* respectively. Each Corel group has 100 textures (768 x 512 pixels) and every texture is divided into 6 subimages, the total number of textures is $6x100 = 600$ for each Corel dataset.

[1]Corel data are distributed through http://www.emsps.com/photocd/corelcds.htm

We have used the Recall measure [14] to evaluate the performance of the retrieval and the precision-recall curves. The results have been computed by using all the images in each dataset as query images. In the ideal case of the retrieval, the top 6 retrieved images would be from the same original subsampled image.

Dataset	PTD	MPEG-7	$LBP_{8,1}$*RGB*
Corel	**81.89%**	67.33%	61.89%
Corel2	**85.78%**	76.11%	72.5 %
CorelV1	**93.44%**	85.94%	77.53%
CorelV2	**93.72%**	88.53%	81.47%
CorelTex	**93.78%**	93.69%	86.89%
CorelSan	**83.61%**	72%	55.54%
all Corel	**86.08%**	66.17%	68.64%

Table I
AVERAGE RETRIEVAL RATE

To compare efficiency, in table I we show retrieval rates for the 6 datasets using our *PTD* and two different descriptors that combine color and texture in different ways. The descriptors used in the comparison are the standard MPEG-7 descriptors [15] HTD and SCD as they are combined in [2], and the extension of LBP descriptor to color proposed in [3]. The Average Retrieval rate computed shows how our *PTD* overcomes both the $LBP_{8,1}$ *RGB* and the MPEG-7 descriptors for the six Corel datasets and also for the dataset that is the union of the 6 Corel datasets. We also show the precision-recall curves using the same datasets in Fig.3 that confirms the previous results over the precision range.

IV. Conclusions

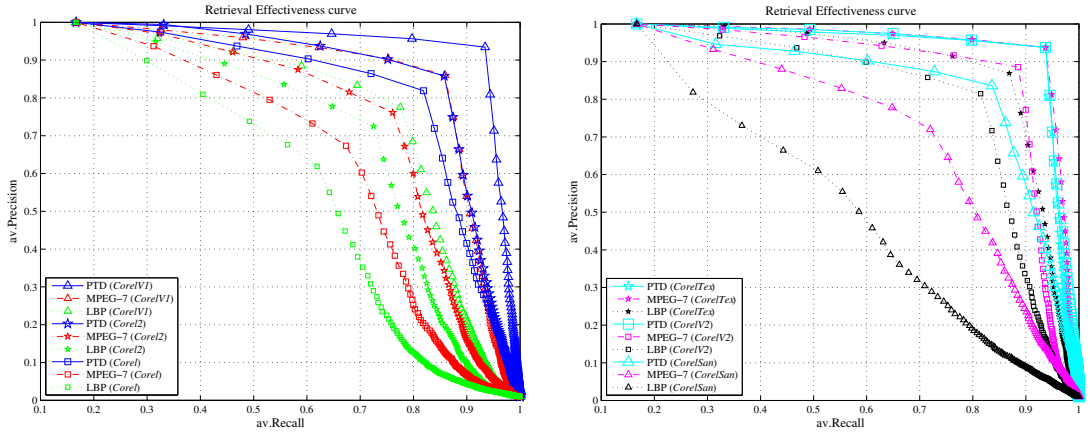
In this paper we propose the *Perceptual Texton Descriptor (PTD)* that implements the original definition of texton given in the Julesz's Texton theory. Combination of color and shape attributes is done through the definition of the p-blob concept. The descriptor is defined on two different texton spaces with some interesting properties. Spaces are of low dimension implying low-redundancy, they are based on axes that correlate with perceptual properties that are bounded by their own perceptual nature avoiding normalizations between cues. This perceptual correlation allows to build a vocabulary of visual words with a semantic with basic linguistic terms. Finally, we show promising results on a retrieval experiment on a texture dataset that present a high degree of diversity.

Acknowledgment

This work has been partially supported by projects TIN2007-64577 and Consolider-Ingenio 2010 CDS2007-35100018 of Spanish MEC (Ministry of Science).

References

[1] Y. Chun, N. Kim, and I. Jang, "Content-based image retrieval using multiresolution color and texture features," *IEEE Trans. on Multimedia*, vol. 10, no. 6, pp. 1073–1084, October 2008.



(a) (b)

Figure 3. Precision-Recall curves of *PTD*, *MPEG7 (HTD and SCD)* and $LBP_{8,1}RGB$ for different datasets. (a) *Corel*, *CorelV1* and *Corel2* datasets. (b) *CorelV2*, *CorelTex* and *CorelSan* datasets.

[2] R. Dorairaj and K. Namuduri, "Compact combination of MPEG-7 color and texture descriptors for image retrieval," in *Conference on Signals, Systems and Computers. Conference Record of the Thirty-Eighth Asilomar*, vol. 1, 2004, pp. 387–391.

[3] T. Mäenpää and M. Pietikäinen, "Classification with color and texture: jointly or separately?" *Pattern Recognition*, vol. 37, no. 8, pp. 1629–1640, 2004.

[4] H. Yu, M. Li, H. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," in *International Conference on Image Processing*, 2003, pp. 24–28.

[5] Y. Zhong and A. K. Jain, "Object localization using color, texture and shape," *Pattern Recognition*, vol. 33, no. 4, pp. 671–684, 2000.

[6] B. Julesz and J. Bergen, "Textons, the fundamental elements in preattentive vision and perception of textures," *Bell Systems Technological Journal*, vol. 62, no. 6, pp. 1619–1645, 1983.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Nineth IEEE International Conference on Computer Vision and Pattern Recognition (ICCV 2003)*, 2003, pp. 1470–1477.

[8] T. Serre, L.Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, 2007.

[9] G. Burghouts and J. Geusebroek, "Material-specific adaptation of color invariant features," *Pattern Recogn. Lett.*, vol. 30, no. 3, pp. 306–313, 2009.

[10] P. Quelhas and J. Odobez, "Natural scene image modeling using color and texture visterms," in *International Conference on Image and Video Retrieval (CIVR)*, 2006.

[11] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[12] M. Varma and A. Zisserman, "Texture classification: Are filter banks necessary?" *IEEE Conference on Computer Vision and Pattern Recognition (CCVPR)*, vol. 2, pp. 691–698, 2003.

[13] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, June 2007.

[14] J. Smith, "Image retrieval evaluation," in *Proc. IEEE Workshop on Content - Based Access of Image and Video Libraries*. Los Alamitos, CA, USA: IEEE Computer Society, 1998, pp. 112–113.

[15] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7*. John Wiley & Sons, 2003.