What are Textons?

SONG-CHUN ZHU, CHENG-EN GUO, YIZHOU WANG AND ZIJIAN XU
Departments of Statistics and Computer Science, University of California, Los Angeles,
Los Angeles, CA 90095, USA
sczhu@stat.ucla.edu
cguo@stat.ucla.edu
wangyz@stat.ucla.edu
zjxu@stat.ucla.edu

Received December 10, 2002; Revised August 27, 2003; Accepted October 3, 2003
First online version published in November, 2004

Abstract. Textons refer to fundamental micro-structures in natural images (and videos) and are considered as
the atoms of pre-attentive human visual perception (Julesz, 1981). Unfortunately, the word “texton” remains a
vague concept in the literature for lack of a good mathematical model. In this article, we first present a three-level
generative image model for learning textons from texture images. In this model, an image is a superposition of
a number of image bases selected from an over-complete dictionary including various Gabor and Laplacian of
Gaussian functions at various locations, scales, and orientations. These image bases are, in turn, generated by a
smaller number of texton elements, selected from a dictionary of textons. By analogy to the waveform-phoneme-
word hierarchy in speech, the pixel-base-texton hierarchy presents an increasingly abstract visual description
and leads to dimension reduction and variable decoupling. By fitting the generative model to observed images, we can
learn the texton dictionary as parameters of the generative model. Then the paper proceeds to study the
geometric, dynamic, and photometric structures of the texton representation by further extending the generative model
to account for motion and illumination variations. (1) For the geometric structures, a texton consists of a number of
image bases with deformable spatial configurations. The geometric structures are learned from static texture images.
(2) For the dynamic structures, the motion of a texton is characterized by a Markov chain model in time which
sometimes can switch geometric configurations during the movement. We call the moving textons as “motons”. The
dynamic models are learned using the trajectories of the textons inferred from video sequence. (3) For photometric
structures, a texton represents the set of images of a 3D surface element under varying illuminations and is called
a “lighton” in this paper. We adopt an illumination-cone representation where a lighton is a texton triplet. For a
given light source, a lighton image is generated as a linear sum of the three texton bases. We present a sequence
of experiments for learning the geometric, dynamic, and photometric structures from images and videos, and we
also present some comparison studies with K-mean clustering, sparse coding, independent component analysis, and
transformed component analysis. We shall discuss how general textons can be learned from generic natural images.

Keywords: textons, motons, lightons, transformed component analysis, textures

1. Introduction

The purpose of vision, biologic and machine, is to compute a hierarchy of increasingly abstract interpretations
of the observed images (or image sequences). Therefore it is of fundamental importance to know what
are the descriptions used at each level of interpretation. By analogy to physics concepts, we wonder what
are the visual “electrons”, visual “atoms”, and visual “molecules” for visual perception. The pursuit of basic
images and perceptual elements is not just for intellectual curiosity but has important implications in a series of practical problems. For example,

1. **Dimension reduction.** Decomposing an image into its constituent components reduces information redundancy and leads to lower dimensional representations. As we will show in later examples, an image of $256 \times 256$ pixels can be represented by about 500 image bases, which are, in turn, reduced to 50–80 texton elements. The dimension of representation is thus reduced by about 100 folds. Further reductions are achieved in motion sequences and lighting models.

2. **Variable decoupling.** The decomposed image elements become more and more independent of each other and thus are spatially nearly decoupled. This facilitates image modeling which is necessary for visual tasks such as segmentation and recognition.

3. **Biologic modeling.** Micro-structures in natural images provide ecological clues for understanding the functions of neurons in early stages of biologic vision systems (Barlow, 1961; Olshausen and Field, 1997).

In the literature, there are several threads of research investigating fundamental image structures from different perspectives, with many questions left unanswered.

Firstly, in **neurophysiology**, the cells in the early visual pathway (retina, LGN, and V1) of primates are found to compute some basic image structures at various scales and orientations (Hubel and Wiesel, 1962). This motivated some well-celebrated image pyramid representations including Laplacian of Gaussians (LoG), Gabor functions, and their variants (Daugman, 1985; Simoncelli et al., 1992). However, very little is known about how V1 cells are grouped into larger structures in higher levels (say, V2 and V4). Similarly, it is unclear what are the generic image representations beyond the image pyramids in image analysis.

Secondly, in **psychophysics**, Julesz (1981) and colleagues discovered that pre-attentive vision is sensitive to some basic image features while ignoring other features. His experiments measured the response time of human subjects in detecting a target element among a number of distractors in the background. For example, Fig. 1 shows two pairs of elements in comparison. The response time for the left pair is instantaneous (100–200 ms) and independent of the number of distractors. In contrast, for the right pair the response time increases linearly with the number of distractors. This discovery was very important in psychophysics and motivated Julesz to conjecture a pre-attentive stage that detects some atomic structures, such as elongated blobs, bars, crosses, and terminators (Julesz, 1981), which he called “textons” for the first time.

The early texton studies were limited by their exclusive focus on artificial texture patterns instead of natural images. It was shown that the perceptual textons could be adapted through training (Karni and Sagi, 1991). Thus the dictionary of textons must be associated with or learned from the ensemble of natural images. Despite the significance of Julesz’s experiments, there have been no rigorous mathematical definitions for textons. Later in this paper, we argue that textons must be defined in the context of a generative model of images.

Thirdly, in **harmonic analysis**, one treats images as 2D functions, then it can be shown that some classes of functionals (such as Sobolev, Hölder, Besov spaces) can be decomposed into bases, for example, Fourier, wavelets (Coifman and Wickerhauser, 1992), and more recently wedgelets and ridgelets (Donoho et al., 1998). It was proven that the Fourier, wavelets,

![Figure 1](image_url)  
**Figure 1.** Two typical examples of searching a target element among a number of background distractors. The search time for the left pair is constant independent of the number of distractors, while it increases linearly with the number of distractors for the right pair. After Julesz (1981).