Pattern Recognition 43 (2010) 2609–2619

Contents lists available at ScienceDirect

Pattern Recognition

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

Total variation, adaptive total variation and nonconvex smoothly clipped absolute deviation penalty for denoising blocky images

Aditya Chopra [a], Heng Lian [b,*]

[a] School of Computing Sciences, VIT University, Vellore, TN, India
[b] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 19 August 2009
Received in revised form
23 March 2010
Accepted 25 March 2010

Keywords:
MM algorithm
SCAD penalty
Total variation denoising

ABSTRACT

The total variation-based image denoising model has been generalized and extended in numerous ways, improving its performance in different contexts. We propose a new penalty function motivated by the recent progress in the statistical literature on high-dimensional variable selection. Using a particular instantiation of the majorization-minimization algorithm, the optimization problem can be efficiently solved and the computational procedure realized is similar to the spatially adaptive total variation model. Our two-pixel image model shows theoretically that the new penalty function solves the bias problem inherent in the total variation model. The superior performance of the new penalty function is demonstrated through several experiments. Our investigation is limited to "blocky" images which have small total variation.

1. Introduction

Denoising is probably the most common and most studied problem in image processing. Approaches developed so far include many methods arising from the field of engineering, computer science, statistics and applied mathematics. There are several popular classes of existing denoising algorithms, from simple linear neighborhood filtering to mathematically more involved wavelet methods based on solid statistical foundations [1–3]. The PDE-based methods first proposed in [4] are unique in their formulation of images as functions in a suitable function space. Relatively few comparison studies exist among different methods, which is quite understandable due to (i) there are a large number of existing denoising approaches with many different modifications and extensions; (ii) the success or failure of different approaches depends largely on the characteristics exhibited by different types of images, whether cartoon or natural scene images, grayscale or colored, textured or solid objects. One exception is the work [5] which compared the standard total variation (TV) model with wavelet denoising and finds that TV is inferior for some standard test images. With different fine tunings and extensions available in both the class of PDE-based and wavelet-based methods, such as using higher order derivatives or correlated wavelet coefficients, it is still hard to judge from their results the relative merits of these two approaches, although it seems to be the prevailing mindset that the wavelet-based methods work better for general images.

Denoting the unobserved original noiseless image by u, the goal of denoising is to recover this original image given an observed noisy image $f=u+n$, where n denotes the noise. In traditional filtering as well as wavelet-based approaches, we either think of images as $m\times l$ matrices or $N=ml$-dimensional vectors, while the PDE-based method will generally treat images as bivariate functions defined on the unit square $\Omega=[0,1]\times[0,1]$. Introduced in [4], the standard total variation (TV) image denoising method estimates the original image by solving the following minimization problem

$$\hat{u}=\arg\min_u \|f-u\|^2+\lambda TV(u), \tag{1}$$

where $\|.\|$ is the L_2 norm of the function and $TV(u)=\int_\Omega|\nabla u|$ is the total variation norm of u [4]. The regularization parameter λ controls the tradeoff between the fidelity to the observed image and smoothness of the recovered image. Actually the paper [4] used the somewhat equivalent formulation of minimizing the total variation with constraints on the noise level, which is assumed to be known. But the penalized L_2 version stated above is more convenient when the level of the noise is unknown and we will adopt this formulation in our study. Both practically and theoretically, this model is the best understood one among PDE-based methods as of today, where the images are considered as belonging to the space of functions of bounded variation (BV) and the existence and uniqueness of solution is well-established [6–8]. Discrete version of the TV model is considered in [5], arguing that all approaches have to go through the discretization

* Corresponding author. Tel.: +65 65137175.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.022

procedure when implemented anyway. Our point of view is that using either the continuous or discrete formulation for the PDE-based methods makes little difference in practice.

Although the standard TV model above might not be competitive for general image denoising tasks, it is believed to be ideal for blocky images, i.e., images that are nearly piece-wise constant. From a statistical point of view, this can be simply seen by the fact that it penalizes the first order partial derivatives (or, in discrete version, first order differences) and thus shrinks them towards zero. Such images are interesting for at least two reasons. First, examples of blocky images abound in real life, such as vehicle registration plates, traffic signs, postal code on envelopes, etc. A more complicated example in medical imaging is found in [9]. Second, studying of such relatively simple images can usually lead to deeper insights into different denoising approaches. [10] noted the inherent bias in the TV model and proposed the spatially adaptive total variation (SATV) model that applies less smoothing near significant edges by utilizing a spatially varying weight function that is inversely proportional to the magnitude of image derivatives. SATV is a two-step procedure where the weight function obtained from the first step using standard TV is then used to guide smoothing in the second step. The authors showed that with a modest increase in computation, SATV is superior to standard TV in restoring piece-wise constant image features.

Curiously, there is an almost parallel development in the statistical literature in the context of high-dimensional linear regression with variable selection. As explained in the next section, these studies focus on the regression problem where although there exist numerous covariates a priori, most of the regression coefficients are exactly zero, implying that the corresponding covariates have no effects on the response variable. Thus shrinking most regression coefficients to zero is a viable strategy for efficient estimation. For piece-wise constant images, with first derivatives in most locations exactly equal to zero, shrinking them to zero is thus also a reasonable approach. Taking advantage of this observation, we propose to adapt the smoothly clipped absolute deviation (SCAD) penalty [11,12] that has become extremely popular in the statistical community for our image denoising task. Although in the case of TV model the correspondence between the functional-analytical approach and the statistical approach seems to be well-known, and some have studied in detail the properties of total variation from a statistical point of view [13,14], these statistical works are only restricted to the one-dimensional case. Besides, as far as we know, the parallelism stated above has not been fully utilized and in particular the SCAD penalty has not been applied to penalize the first order differences even in the one-dimensional case. Besides its superior performance in practice, there are several advantages of SCAD penalty compared to SATV, most notably getting rid of the extra parameter that a user needs to tune for SATV in implementation. As mentioned before, we think using either discrete or continuous formulation formally makes little difference, but we choose to use the continuous formulation since it can simplify description and notation significantly. The only problem is that the objective functional using the SCAD penalty being nonconvex, existence of solution is not guaranteed. The theoretically inclined reader might want to think in discrete terms so that such technical point does not arise. Our computational experiments show that SCAD is superior to SATV in terms of mean squared error (MSE). Although MSE is notorious for describing the visual quality of an image, it is arguably less so for blocky images where MSE can describe the accuracy of restoration rather faithfully.

The rest of the paper is organized as follows. In the next section, we briefly review the TV and the SATV model and point out the almost trivial connection to Lasso and the adaptive Lasso developed in the statistical literature so that we hope readers from both fields can follow the motivation and development of the current paper. In Section 3, we adapt the SCAD penalty for our image denoising problem and discuss some properties in detail in this context. We also developed a majorization-minimization procedure using first order Taylor expansion so that the computation involved simply reduces to that similar to the SATV model, although with a different weight function. In Section 4, we will briefly review a method called Monte-Carlo SURE [15] for regularization parameter selection which is used in our study when required. In Section 5, several computational experiments are used to show the superiority of the proposed method in denoising blocky images. In these experiments, we also intentionally emphasize the difficulty encountered with SATV model in tuning its performance. We conclude the paper with a discussion in Section 6.

2. Review of the TV and SATV model

The TV model proposed by [4] and presented above in Eq. (1) has received a great deal of attention in the last decade. In [10], the authors argued that it is desirable that less smoothing is carried out where there is more detail in the image. This motivated the replacement of TV norm by the following more general weighted TV functional

$$TV_w(u) = \int_\Omega w(x,y)|\nabla u(x,y)|\,dx\,dy. \tag{2}$$

The weight w should be small in the presence of an edge so that less smoothing is performed near an edge. [10] used a weight function inversely proportional to the partial derivatives, with a parameter $e>0$ added both to avoid dividing by zero and to be used as a tuning parameter to control the amount of adaptivity. Thus in their proposal of the spatially adaptive total variation (SATV) model $w=1/(|u_x|+e)+1/(|u_y|+e)$ where u_x and u_y are the partial derivatives. [10] used a two-step method. In the first step the standard TV model (1) is used to estimate u based on which the partial derivatives (first order differences) are computed. Then the derivatives are used in (2) to compute the final restored image. If e is chosen sufficiently large, SATV basically reduces to the standard TV. On the other hand, if e is too small, artificial edges will appear and the algorithm will be numerically unstable as well. We will see in our simulations that the result is somewhat sensitive to the choice of e and the appropriate amount of adaptivity is not universal to all images, which makes it difficult to choose e in practice, or leads to a sizable increase on the amount of computation required to say the least.

As we mentioned in the introduction, there is an almost parallel line of development in the statistical literature that uses the same idea of SATV in a different context. Consider a linear regression problem $y_i=\mathbf{x}_i^T\beta+\varepsilon_i$ based on independent and identically distributed (i.i.d.) data $\{y_i,\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i=(x_{i\,1},\ldots,x_{ip})^T$ are the covariates, $\beta=(\beta_1,\ldots,\beta_p)^T$ are the regression coefficients, and ε_i is a zero mean noise. Sometimes one has good reasons to believe that only a few of the x_{iq}'s are related to y_i, i.e., many of the β_q's are exactly zero. In these situations it is desirable to design an approach that shrinks many regression coefficients to zero automatically. Lasso [16] does exactly that and IS formulated as the minimization of the following objective function:

$$\sum_{i=1}^n \|y_i-\mathbf{x}_i^T\beta\|^2+\lambda\sum_{i=1}^p|\beta_i|.$$

It is now well-known that this algorithm encourages many coefficients to be exactly zero as desired due to the use of L_1