Image denoising with k-nearest neighbor and support vector regression

Bram van Ginneken & Adriënne Mendrik
Image Sciences Institute, University Medical Center Utrecht, the Netherlands

Abstract

Denoising is an important application of image processing, especially for medical image data. These images tend to be very noisy when a low radiation dose, less harmful to the patient, is used for acquisition. For computed tomography (CT) data, it is possible to simulate realistic low dose images from the raw scanner data. We use this data to construct a supervised denoising system, that learns an optimal mapping from input features to denoised voxel values. As input features we use several general filters and the output of existing standard noise reduction filters, notably non-linear diffusion schemes. After feature selection, these are mapped to the denoised values by k-nearest neighbor and support vector regression. The resulting regression denoising systems are shown to perform significantly better than non-linear diffusion schemes, Gaussian smoothing and median filtering in experiments on CT chest scans.

1. Introduction & Outline

Computed tomography (CT) is the workhorse of modern radiology. The latest multi-slice CT scanners offer a unique opportunity to visualize the inside of the human body with very high spatial and temporal resolution. The major disadvantage of CT is its use of ionizing radiation, which induces a risk of developing cancer. Therefore low dose and ultra low dose scanning are becoming standard practice in radiology, especially for mass screening, repeated examinations and imaging of young patients. These low-dose images are very noisy, as can be seen in Fig. 1(b). Consequently, the use of advanced techniques for noise removal is extremely important for the analysis of CT scans.

The scientific literature offers an enormous variety of techniques for image denoising. They include non-linear diffusion, neighborhood filters, median filtering, wavelet schemes and other transform filters, Wiener filtering, total variance minimization, and many more [3]. What these techniques have in common is that, although they are based on certain assumptions (a model) of both the noise and the image, they are *not* supervised. That is, they are not tuned for a specific application by learning, from provided examples, the mapping from noisy images to high quality denoised images. These 'classical' noise reduction algorithms typically contain a number of parameters that need to be optimized for a particular application. This issue tends to be largely ignored in the literature. We have observed that in practice, for denoising medical images, optimal settings for parameters vary per application, and even per image and within images. We also found that for the purpose of denoising CT data, different algorithms tend to be superior in different parts of one particular input scan. This finding suggests that denoising may be improved by combining the output of different noise reduction algorithms using supervised techniques. In this work it will be shown that such an improvement can actually be achieved with regression. We perform regression on a voxel level. For every voxel we take as input the noisy data and the result of applying a number of filters and several denoising algorithms with different parameter settings to that data. From this input we infer the output, the denoised voxel value, through regression. Two regression techniques are compared: k-nearest neighbor and support vector machines.

In order to train the regression, examples are required. Often it is hard to obtain such data. In the case of CT scans, one may scan a patient twice, with a standard (high) clinical dose and with a much lower dose. However, in that case the two scans are not aligned precisely enough to be able to compare voxels from the low and high dose scan directly. It is possible, however, to simulate low dose data from a high dose scan by adding physically realistic noise to the raw CT scanner data before reconstruction [1]. Data sets constructed in this way are use in this work.

Supervised noise reduction has not received much attention, certainly not within medical image analysis. One exception is the work of Kwok and Tsang [6], who use kernel principal component analysis to build models of handwritten digits and show how to obtain the image before projection from the principal components. With this technique extremely noisy versions of these digits can be restored very effectively. This technique is difficult to apply, however, to general noise reduction for images that may contain an endless variety of structures at any position. Another dif-

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00

ference with the technique proposed here is that the noise is not modeled explicitly.

2. Method

Regression is defined as computing the mapping from n input values (features) to a single continuous target value. The mapping is determined from a set of training examples. Linear regression takes a linear combination of the features (where the weights are typically found by minimizing the sum of squared errors on the training data) and was found to be too restrictive for the task considered here. k-nearest neighbor regression (kNNR) is a non-parametric technique that finds for a given input the k nearest neighbors among the training examples in feature space and returns the average output of these neighbors. In this work the Euclidean distance metric was used after scaling each feature to unit variance on the training data. A fast tree-based algorithm was used to compute the nearest neighbors [2]. The single free parameter in kNNR is k. Support vector machines, commonly used as classifiers, can also be used for regression. The most common approach is to adopt a linear ϵ-insensitive loss function. The implementation in LIBSVM [5] was used, with Gaussian functions as kernel and all features scaled to unit variance. This leads to a support vector regression (SVR) system with three free parameters, ϵ, the penalty term C and the scale σ of the Gaussian.

Two types of features are investigated: a general set of features that describe image structure based on Gaussian derivative filters and the output of several well-known noise reduction schemes obtained with different parameter settings.

The rationale for using these image structure descriptors is that they encode what an image locally looks like in a way similar to a Taylor expansion of a function, and thus provide neighborhood information to the denoising algorithm. We compute the output of Gaussian filters and Gaussian derivative filters of order 1 (L_x, L_y, L_z) and 2 (L_{xx}, L_{xy}, L_{xz}, L_{yy}, L_{yz}, L_{zz}) at scales $\sigma = 2^i$ with $i = -1, \ldots, 4$. Derivatives are combined into several rotationally invariant filters: the gradient magnitude ($\sqrt{L_x^2 + L_y^2 + L_z^2}$) and the Laplacian ($L_{xx} + L_{yy} + L_{zz}$).

The use of outputs of different noise reduction schemes is motivated by the observation that the best performing noise filter (type and parameters) appears to vary from region to region and image to image. Providing different estimates of the denoised voxel value as input features could be interpreted as using the regression system as a filter combining and parameter selection system.

The Gaussian filter applied at different scales can be considered one type of noise reduction scheme. The second scheme that was used is median filtering in a $n \times n \times n$ neighborhood with $n = 3, 5, 7$.

The third scheme is Perona-Malik filtering (PM) which is a non-linear inhomogeneous isotropic diffusion filter [7]. The idea is that close to edges, characterized by high image gradients, the Gaussian neighborhood over which one averages should be smaller than in homogeneous regions. The regularized form of PM proposed in [4] was used. In this scheme there are four parameters: a time step t which should be sufficiently small to ensure numerical stability and which is fixed to 0.15; the scale for computing the gradient that controls the size of the neighborhood over which to average, fixed to 1 voxel; the threshold λ which determines what should be considered a 'high gradient' and controls the amount of 'non-linearity' of the filter; and the number of iterations T for which the iterative scheme is applied where a high T leads to more smoothing. A priori it is not clear how to choose λ and T. We use $\lambda = 20$ and 50 and $T = 1, 2, \ldots, 20$. These ranges were determined in pilot experiments.

The fourth and final method is edge enhancing diffusion (EED), proposed in [8]. This is an anisotropic diffusion scheme where the kernel over which to smooth is oriented to lie *along* edges instead of across them, which leads to increased steepness of edges while blurring homogeneous regions. EED has the same parameters as PM, and λ is again set to 20 and 50 and T to $1, 2, \ldots, 10$.

Clearly many other noise reduction schemes could have been chosen. Our choice for non-linear diffusion schemes was motivated by their popularity in medical image analysis. The framework we propose can obviously use any other type of feature.

Altogether this results in a large number of features: the input image itself, 6 Gaussian blurred versions, 54 derivatives, 12 higher order invariants, 3 median filters, 40 PM filters and 20 EED filters = 136 features in total. Computing all these features for actual denoising of large volumetric medical image data sets would require prohibitive amounts of computation time and memory. Therefore feature selection was used in the experiments with sequential forward selection (SFS) [9].

3. Experiments & Results

Experiments were performed on CT chest data sets of four different patients with a voxel resolution of $0.7 \times 0.7 \times 0.7$ mm. The scans were acquired with a standard clinical dose of either 150 or 130 mAs. These images are referred to as 'high dose' data in the remainder of the paper. Using a noise simulation package [1], physically realistic low dose scans of 15 mAs were constructed. In each scan, a number of non-overlapping volumes of interest (VOIs) were manually delineated in different areas of the scan (lungs, heart, spine, muscle, and so on). The total number of VOIs is

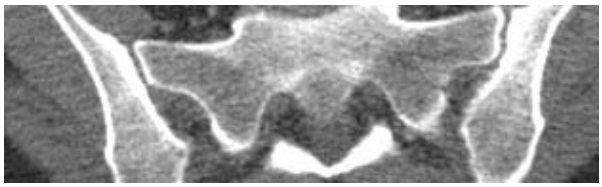
58, average volume 46 cm^3. All experimental results are obtained with cross-validation, with data subdivided in two folds, each with VOIs from two patients. Training, including feature selection and determination of the optimal parameter settings for the regression systems, was performed using only data from one fold, the resulting system was used on the other fold for testing.

For the sequential forward feature selection only kNNR was used. Feature selection with SVR was omitted because SVR training and testing is orders of magnitude slower and SVR performance is very sensitive to proper choices for the internal parameters C and σ which resulted in impractically long training times. kNNR performance, on the other hand, is relatively insensitive to the choice of k. For feature selection, a training set of 130,000 randomly picked voxels from the VOIs in the training fold was used, and another set of 6,000 randomly picked voxels (from the same training fold) was used to evaluate performance of a set of features. This procedure produces subsets of n features. To determine the optimal n, a third set of 6,000 randomly picked voxels (from the same training fold) was used. This third set was also used to determine the optimal k for kNNR and, by hill-climbing, C and σ for SVR. In pilot experiments it was observed that the setting for ϵ was not very critical so this parameter was fixed to 0.1. For training the SVR, a set with less voxels, 13,000, was used to keep training times acceptable (in the order of 20 hours for parameter optimization on a standard PC).

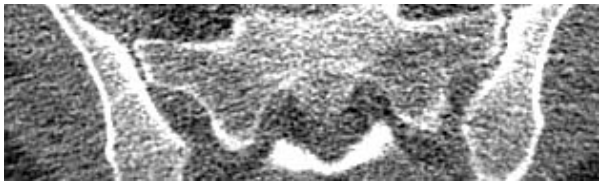
The optimal number of selected features was four and five, for both training folds respectively. Selected features were PM filters (four times) and EED filters (five times), all with different settings. Interestingly, derivatives and image invariants were never selected among the first 10 features.

To compare performance between schemes, the mean absolute distance (MAD) between the inferred and the real high dose data was computed, averaged per VOI. Table 1 lists the mean and standard deviation over all VOIs for the unprocessed low dose data, the results obtained with the best parameter settings for blurring, median filtering, PM and EED and for kNNR and SVR. The table also lists p values from a paired two-tailed t-test between any pair of filters. The result is that the regression schemes are significantly better than all other schemes. Of the latter, PM performs best. There is no significant difference between kNNR and SVR.

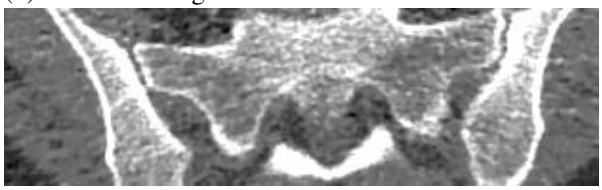
Figure 1 shows an example result.



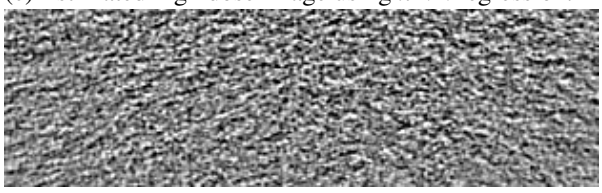
(a) Original, high dose image.



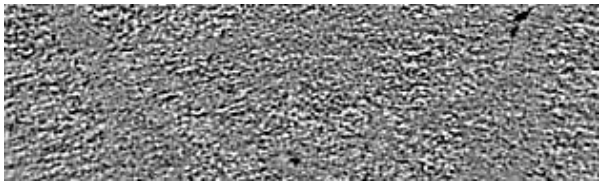
(b) Low dose image.



(c) Estimated high dose image using kNN regression.



(d) True noise image: (b) subtracted from (a).



(e) Estimated noise image: (c) subtracted from (a).

Figure 1. Example results of denoising with kNN regression

4. Discussion

It has been shown that by providing example data of noisy and high quality scans, the performance of standard noise reduction techniques can be improved through the use of regression. The feature selection procedure gives insight in which combination of features is beneficial for this task. It turned out that all selected features were outputs from other noise reduction schemes, and the addition of general image structure filters did not improve performance. It is possible to construct a decent noise removal regression system using only filters derived from Gaussian derivatives as input, but such a system does not outperform standard nonlinear diffusion schemes. This may indicate that the underlying models for PM and EED are appropriate for noisy medical image data.

Note that the proposed technique can be used with any other feature. Thus it may be beneficial to add the output of other noise reduction schemes to the feature set. As long as

Table 1. Results of denoising in terms of mean absolute distance (MAD) to high dose truth with standard deviation in parentheses. Values are in Hounsfield units. P values for a paired two-tailed t-test are listed for any pair of schemes. For each scheme, best results for all parameters settings were selected. LD = low dose scan (no processing). GB = Gaussian blur, $\sigma = 2.0$. M = median filtering $3 \times 3 \times 3$ neighborhood. PM = Perona-Malik, $\lambda = 50, T = 5$. EED = Edge enhancing diffusion, $\lambda = 20, T = 3$. kNNR = k-nearest neighbor regression. SVR = support vector regression.

	LD	GB	M	PM	EED	kNNR	SVR
mean MAD	37.42 (14.09)	24.63 (7.68)	24.99 (7.27)	20.13 (6.69)	20.12 (6.92)	19.05 (6.80)	19.15 (6.71)
LD		1.1×10^{-11}	1.0×10^{-11}	1.7×10^{-19}	5.7×10^{-20}	6.4×10^{-21}	3.2×10^{-21}
GB			1.0×10^{-1}	3.1×10^{-14}	2.8×10^{-14}	2.9×10^{-16}	3.7×10^{-15}
M				2.8×10^{-20}	1.8×10^{-18}	6.5×10^{-21}	3.9×10^{-20}
PM					9.5×10^{-1}	2.0×10^{-6}	2.0×10^{-6}
EED						1.4×10^{-7}	2.4×10^{-5}
kNNR							4.3×10^{-1}

there is not a single denoising filter that consistently outperforms other systems, a combination through regression may further improve results. The proposed system can also be thought of as a way to automatically choose a complex combination of different filters and different settings for these filters, optimally tuned for a particular task. This is of great practical value.

The performance measure, mean absolute difference, is not necessarily a good indicator of filter quality. Many denoising filters suffer from the fact that they remove small image details, which may actually be very important for a radiologist in this particular task. The same is true for the system presented here, as can be seen from Fig. 1. The diffusion schemes used as input also tend to smooth too much at their optimal settings (in terms of MAD). Buades et al. [3] suggest to look at the estimated noise image, which is also shown in Fig. 1. It looks similar to the real noise but it still contains certain structural information not present in the real noise. Despite the limitations of MAD, it is a good measure for comparing different schemes. A practical evaluation would investigate the use of denoised images in a clinical setting, such as for detection or quantification of abnormalities.

Upon visual inspection of Fig. 1, there seems definitively room for improvement. In the denoised image, noisy structures have remained at some positions. At the same time, the texture in the denoised image is also not realistic, it is smoother than in the real high dose data. Denoising remains a very challenging problem.

The results of kNNR and SVR were not significantly different ($p = 0.43$). Note however that SVR uses a factor of 10 less input data. If the kNNR system is trained with only that amount of data it performs substantially worse. Moreover, the feature selection was done with kNNR and may thus not be optimal for SVR. This may indicate that better results with SVR are possible. Run-time, both for testing and training, is much slower for SVR, however.

To further improve results it may be useful to investigate the use of more advanced features that are tailored towards the differences in noise and image structure for this particular application. In addition, iterative regression schemes for noise and signal estimation may also further improve results at the expense of longer computation times.

References

[1] O. Amir, D. Braunstein, and A. Altman. Dose optimization tool. In *Proceedings of the SPIE*, volume 5029, pages 815–821, 2003.

[2] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[3] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale modeling and simulation*, 4(2):490–530, 2005.

[4] F. Catté, P. L. Lions, J. M. Morel, and T. Coll. Image selective smoothing and edge-detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.

[5] C. Chang and C. Lin. LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] J. T. Kwok and I. W. Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15(6):1517–1525, 2004.

[7] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:629–639, 1990.

[8] J. Weickert. *Anisotropic diffusion in image processing*. B.G. Teubner, Stuttgart, 1998.

[9] A. Whitney. A direct method of nonparamteric measurement selection. *IEEE Transactions on Computers*, 20:1100–1103, 1971.