Neurocomputing 74 (2011) 1402–1410

Contents lists available at ScienceDirect

Neurocomputing

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

Mode estimation in high-dimensional spaces with flat-top kernels: Application to image denoising

Arnaud De Decker [a,*,1], Damien François [a], Michel Verleysen [a], John A. Lee [a,b,2]

[a] Machine Learning Group, Université catholique de Louvain, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
[b] Molecular Imaging and Experimental Radiotherapy, Université catholique de Louvain, Avenue Hippocrate 55/5469, B-1200 Brussels, Belgium

ARTICLE INFO

Available online 21 February 2011

Keywords:
Mode estimation
Curse of dimensionality
Concentration of norms and distances
Similarity kernel
Image denoising

ABSTRACT

Mode estimation is extensively studied in statistics. One of the most widely used methods of mode estimation is hill-climbing on a kernel density estimator with gradient ascent or a fixed-point approach. Within this framework, Gaussian kernels proves to be a natural and intuitive option for non-parametric density estimation. This paper shows that in the case of high-dimensional data, mode estimation can be improved by using differently shaped kernels, called flat-top kernels. The improvement are illustrated with an image denoising application, in which pictures are decomposed into small patches, i.e. groups of adjacent pixels, that are vectorized. Noise in the patches can be attenuated by substituting them with the closest mode in the observed distribution of patches. The quality of the denoised picture then depends on the accuracy of mode estimation in a high-dimensional space. Experiments conducted on usual benchmarks in the image processing community show that flat-top kernels outperform the Gaussian one.

1. Introduction

Mode estimation is of utmost importance in many domains, and particularly in signal and image processing. These last two fields have a great place in our every day life, as widely used devices in multimedia, entertainment and professional applications in medicine, geography, or security use advanced signal processing and image denoising techniques.

One of the most striking use of mode estimation methods is image denoising. Noise in pictures can arise because of poor light condition, short exposure and low photon detection, among others. The origin of this noise determines its statistical properties; it can be either additive or multiplicative, Gaussian, Poissonian, or follow a more complex model.

Some of the most studied denoising methods have been developed in the field of mode estimation [1–4] and robust statistics [5,6]. The underlying assumption is that the noisefree data should consist of a few repeated patterns, at least locally or temporarily. From a statistical point of view, the data distribution has therefore several modes, whose width depends on the noise standard deviation. Hence, under usual noise assumptions, the top of any mode indicates a location that should correspond to a noisefree pattern. In practice, mode estimation is achieved by running a hill-climbing procedure [2,3] on a kernel density estimator (KDE) [1]. This paradigm drives many filters in image processing. The mean-shift [2,3], local M-smoothers [7,8] and bilateral filtering [9,10] are the most known among these filters.

Denoising by mode estimation shows its full power when it is applied on multidimensional data. As in classification tasks, adding dimensions is thought of as a mean to increase the gap between the modes, and therefore the probability to identify them correctly. In most publications, hill-climbing on multivariate probability density functions (PDFs) involves a straightforward generalization of Parzen's window estimator [1,11]. Gaussian kernels are used in monodimensional as well as multidimensional spaces. In the case of image filtering, instead of single pixels, the filtering process uses blocks of images called *patches*, which are high-dimensional vectors. The non-local means (NLmeans) [12,13], unsupervised information-theoric adaptative filtering (UINTA) [14] and Bayesian approaches [15] are the most popular of these patch-based filters. Adaptive patch sizes [16] and iterative updates [16,17] have also been investigated.

In mode estimation methods, pixels (or patches) are compared using a notion of similarity. This similarity is a function (typically a Gaussian kernel) of the Euclidean distance between the intensity of patches to be compared. The literature contains many

* Corresponding author.
E-mail addresses: arnaud.dedecker@uclouvain.be (A. De Decker), damien.francois@uclouvain.be (D. François), michel.verleysen@uclouvain.be (M. Verleysen), john.lee@uclouvain.be (J.A. Lee).

[1] The author is funded by a Belgian F.R.I.A. grant.
[2] The author is a Research Associate funded by the Belgian National Fund of Scientific Research (F.R.S.-FNRS).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.12.013

publications that give recommendations about the choice of the optimal kernel and bandwidth [18–20] for the one-dimensional case (pixel version of the mode estimation filters). As the patches are high-dimensional vectors, their distance is measured in a high-dimensional space and is subject to the curse of dimensionality [21,22]. The phenomenon of norm and distance concentration [23] is of particular interest, as KDEs involve pairwise distances in radial kernels.

Many other paradigms have been used to remove or attenuate perturbations in order to recover the true image, like wavelets [24–26], anisotropic diffusion [27,28] and partial differential equations [8,29]. Total variation denoising [30–32] uses a regularization method balancing a smoothness and fidelity term to produce a denoising effect. Recently, wavelets and collaborative filtering were combined to produce BM3D [33], a new filtering algorithm that provides extremely competitive denoising performances.

This paper is an extended version of [34]: its goal is to show that using flat-top kernels improves the mode estimation in high-dimensional data, and in particular, improves the filtering results compared to the statistical filters using a Gaussian kernel. Taking the norm concentration into account allows us to define kernels that are shown experimentally to achieve better mode estimation than the usual Gaussian kernel. Visually, these kernels happen to have the shape of a plateau, hence the name 'flat-top' kernel.

The rest of this paper is organized as follows. Section 2 introduces the density estimation using kernels in the multi-dimensional case. Section 3 shows how the hill-climbing mode estimation procedure can be derived from the density estimation, and how image denoising is achieved using this method. Section 4 deals with the counter-intuitive properties of norms and distances in high-dimensional spaces. It also defines similarity kernels that take these properties into account. Section 5 introduces the image filtering algorithms based on mode estimation by hill-climbing the kernel density estimator. Section 6 experimentally compares the denoising performance of patch-based filtering with either a Gaussian kernel or the proposed similarity functions. Finally, Section 7 draws the conclusions.

2. Kernel density estimators

The most widely known KDE is undoubtedly Parzen's window estimator [1]. In the multidimensional case, this non-parametric estimator approximates an unknown PDF $p(\mathbf{x})$ defined on $\mathbb{R}^D$, from which a finite sample denoted by $\mathbf{X}=[\mathbf{x}_i]_{1\le i\le N}$ is drawn. The estimator can be written as

$$\hat{p}(\mathbf{x}_i)=C\sum_{j=1}^{N}\Psi_\sigma(\|\mathbf{x}_i-\mathbf{x}_j\|_2^2), \tag{1}$$

where C is a normalization factor that ensures that $\int\hat{p}(\mathbf{x})\,d\mathbf{x}=1$ and kernel Ψ_σ is a positive and monotonically decreasing function. $\|\cdot\|_2$ denotes the Euclidean norm, and σ is a bandwidth that controls the estimator smoothness. As the argument of Ψ_σ is a norm, the resulting function of $\mathbf{x}$ is a radial kernel. One usually chooses Ψ_σ as proportional to a Gaussian function, that is, $\Psi_\sigma(u)=\exp(-u^2/2\sigma^2)$.

Intuitively, the KDE transforms the discrete PDF of the observed sample, which consists of a finite set of Dirac impulses, into a continuous PDF. For this purpose, each impulse is blurred by replacing it with a weighted, smooth, narrow, and monomodal PDF, such as a Gaussian one.

There exist of course many refinements of Parzen's window estimator. The main differences lie in the type of kernel or in the bandwidth determination [18–20].

3. Mode estimation

As the modes of PDF $p(x)$ correspond to its local maxima, locating the modes of KDE $\hat{p}(\mathbf{x})$ is a way to approximate these maxima. For this purpose, we can run a so-called hill-climbing procedure on $\hat{p}(\mathbf{x})$ [2]. In practice, we can use simple techniques such as a gradient ascent or fixed-point iterations. The different maxima can be reached by changing the initialization point.

Obviously, the KDE smoothness is critical in this process. If it is too low, then the hill-climbing procedure is likely to get stuck in spurious local maxima, leading to insufficient noise reduction. Running the procedure several times with different initializations around the same actual mode of $p(\mathbf{x})$ shows that the mode estimator has a high variance in this case. In contrast, if the KDE is too smooth, then close modes will not be distinguished anymore and their estimates will be biased. Mode seeking usually requires a smoother KDE and thus a larger kernel bandwidth than in PDF approximation tasks. Indeed the accuracy of the mode locations is important in this case, not the discrepancy between the estimated PDF and the true one.

Equating the derivative of Eq. (1) with 0 provides a fixed-point update that is written as

$$\hat{\mathbf{x}}^{(t+1)}=\frac{\sum_{i=1}^{N}\Psi'_\sigma(\|\hat{\mathbf{x}}^{(t)}-\mathbf{x}_i\|_2^2)\mathbf{x}_i}{\sum_{i=1}^{N}\Psi'_\sigma(\|\hat{\mathbf{x}}^{(t)}-\mathbf{x}_i\|_2^2)}, \tag{2}$$

where t is the iteration index. If $\hat{\mathbf{x}}^{(0)}=\mathbf{x}_i$, then iterating the update amounts to climbing toward the closest hill top on the PDF estimate. The same kind of mode estimator can also be derived from robust statistics [5]. Robust statistics aims at finding estimators that are robust against outliers. In this case, this can be achieved by replacing $\|\hat{\mathbf{x}}^{(t)}-\mathbf{x}_i\|_2^2/2\sigma^2$ by upper bounded functions such as Leclerc's $\sigma^2(1-\exp(-\|\hat{\mathbf{x}}^{(t)}-\mathbf{x}_i\|_2^2/2\sigma^2))$. The upper bound limits the influence of the outliers in the mode estimation. The fixed-point maximization of this generalized non-convex estimator leads to a similar update as the hill-climbing procedure on a KDE.

4. Norms, distances, and similarities in high-dimensional spaces

High-dimensional spaces have weird and counter-intuitive properties. The *curse of dimensionality* [35,36] refers to the ensemble of surprising behaviours observed in high-dimensional spaces. One of these phenomena is the so-called *norm concentration*: when increasing the dimensionality of the space, the mean of usual norms and distances increases, but their variance does not change. Thus, in high-dimensional spaces, the relative error made by considering the expectation of the norms of a specific norm value is very small and gets smaller and smaller as the dimension increases: the minimum and maximum distances look also similar. The discrimination power of the norm is thus less important in high-dimensional spaces. For instance, in the case of a D-dimensional zero-mean unit-variance Gaussian distribution, Euclidean norms are χ_D-distributed, as illustrated in Fig. 1: if $\mathbf{x}_1$, $\mathbf{x}_2\sim N(0,1)$ are two independent D-dimensional normal distributions, then the Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is $\sqrt{\|\mathbf{x}_1-\mathbf{x}_2\|^2}$. As the sum of the square of two independent normal variables follows a χ^2 distribution, the Euclidean distance of $\mathbf{x}_1$ and $\mathbf{x}_2$ thus follows a χ distribution.

The aim of a similarity measure is to evaluate how close two vectors are. Intuitively, a similarity measure should be inversely proportional to the distance between the vectors, and two vectors should be considered as similar if they are drawn from the same mode. Let us assume that data $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from a distribution with a single mode located at $\boldsymbol{\mu}_i$. A good similarity

measure between $\mathbf{x}_i$ and $\mathbf{x}_j$ should be high for values of $\|\mathbf{x}_i-\mathbf{x}_j\|_2$ comprised between 0 and the most probable similarity measure in this mode. This measure of similarity can result from the application of a decaying kernel to pairwise distances. Among all the possible kernel choices, the most widely used is the Gaussian function $\psi_\sigma(\mathbf{u})=\exp(-\mathbf{u}^2/2\sigma)$. The application of the Gaussian function maps Euclidean distances within the interval [0,1]. In order to be discriminant, the similarity measure should be close to 1 for distances between 0 and quantile 0.05 of the distance distribution, and close to 0 beyond quantile 0.95.

In the next section, it will be shown that, while the choice of the Gaussian kernel is natural in one-dimensional spaces, it is not efficient in high-dimensional spaces [38]. Fig. 2 (left) illustrates the distance distribution in high-dimensional spaces and some Gaussian kernels with varying σ. It is easy to see that the Gaussian kernel does not have a high discriminant power: it is not possible to find a Gaussian kernel in which most of the similarity decay occurs within the interval given by quantiles 5% and 95% of the distance distribution.

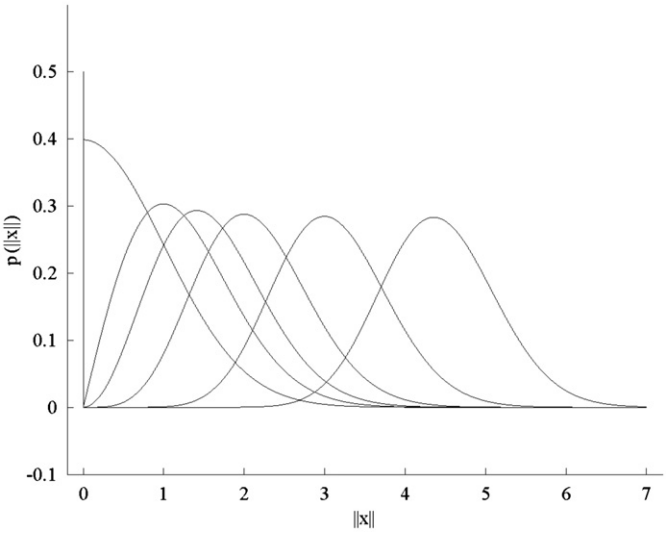

Fig. 1. The Euclidean norm of a random D-dimensional vector $\mathbf{x}$ drawn from a zero-mean identity-covariance normal distribution has a χ_D distribution. $D\in\{1,2,3,5,10,20\}$, from left to right.

In order to find a kernel that satisfies the previous conditions and is discriminant in high-dimensional spaces, let us assume that the data consist of a sample drawn from a mixture of M modes with $1\le j\le M$. Let us define the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ as the probability of observing a larger distance than the one that is measured. That probability is $\mathrm{sim}(\mathbf{x}_i,\mathbf{x}_j)=\mathrm{P}[\sqrt{2}\sigma c\le\|\mathbf{x}_i-\mathbf{x}_j\|_2]$, where $c\sim\chi_D$. The similarity is thus given by the complementary cumulative distribution function (CCDF) of a scaled χ_D variable:

$$\mathrm{sim}(\mathbf{x}_i,\mathbf{x}_j)=\int_{\|\mathbf{x}_i-\mathbf{x}_j\|_2}^{\infty}\frac{\sqrt{2}}{\Gamma(D/2)}\left(\frac{c}{2\sigma}\right)^{D-1}\exp\left(-\frac{c^2}{4\sigma^2}\right)dc=Q\left(\frac{c^2}{4\sigma^2},\frac{D}{2}\right),\tag{3}$$

where Q is the regularized upper incomplete Gamma function [37]. The Gaussian kernel corresponds to the case $D=2$.

Parameter D controls the shape of the kernel; with $D=2$, the kernel is identical to a Gaussian kernel. As D increases, the slope of the kernel gets steeper and moves to the right. With kernel (3), the similarity equals 0.5 when the distance equals the median value of the distribution; depending on how the modes overlap, this may not be the optimal choice. The second parameter σ can then be adjusted in the same way as the bandwidth is adjusted and/or optimized in Gaussian kernels. Kernel (3) is illustrated in Fig. 2 (right); the slope of the kernel function can easily be adjusted to be located in the main part of the distance distribution. This kernel is thus able to satisfy the conditions of a discriminant similarity function in a high-dimensional space.

The CCDF (3) is expensive to compute and it can prove useful to approximate it with the CCDF of an 'all-purpose' distribution, such as the Burr type XII distribution [39]. The scaled CCDF of the Burr type XII distribution is given by $F(b,\lambda,\tau,\theta)=(1-(b/\lambda)^{\tau})^{-\theta}$. A good approximation of the CCDF of the χ_D distribution can be found with $\tau=D$, $\theta=-1$, and $\lambda=\sqrt{2}\sigma$. The Burr CCDF reproduces the shape of the χ_D CCDF, with first a flat top, a steep descent, and a thin tail.

Fig. 3 represents the distance distributions between the patches in the Lena picture (see Fig. 4), for patches of dimension $D=\{1,25,100\}$ (left, central, and right picture respectively. See Section 5 for the definition of the patches). The distance concentration phenomenon is indeed observed: as the dimension increases, the mean of the distance distribution moves away from

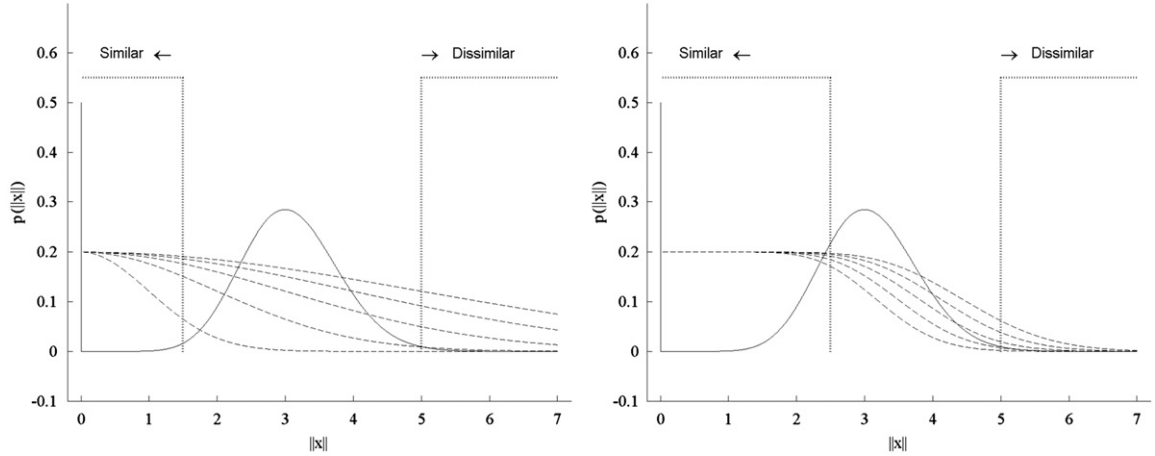

Fig. 2. Pairwise distances between points that are all drawn from the same Gaussian distribution have a distribution given by $p(\|x\|^2)$ (see also Fig. 1). All these points are considered to be similar to each other. A useful similarity measure should thus remain close to one up to main mode and then decrease fast in order to be close to zero in the distribution right tail. Any observed distance in the right tail should be associated with dissimilar points, as its actual probability to occur is otherwise very low. On the left, a Gaussian similarity is obviously too smooth and unable to drop quickly enough between the main mode and the right tail, whatever the bandwidth is (σ). On the right, a flat-top kernel taking $p(\|x\|^2)$ into account behaves more appropriately ($\sigma\in\{1.0,1.1,1.2,1.3,1.4\}$). (Similarity functions are scaled to 0.2 for readability purpose.)

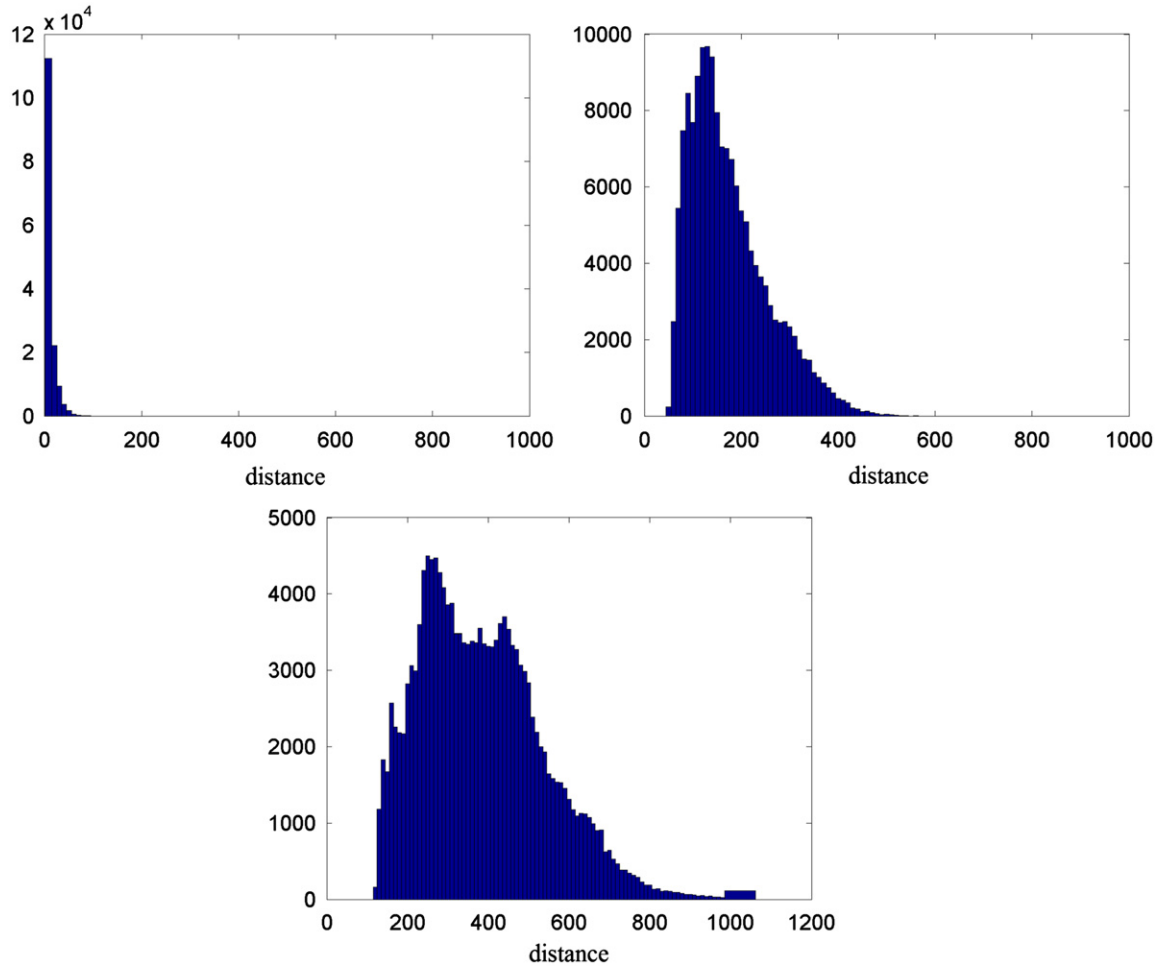

Fig. 3. Pairwise Euclidean distances between patches of Lena (see Fig. 6) picture. The modes of the empirical distance distributions move away from 0 as dimension of the patches increases. Upper row. On the left: one-dimensional patches (single pixels); right: 5×5 patches, $D=25$; lower row: 11×11 patches, $D=121$.

Fig. 4. Original images. Upper row, left: Couple, middle: Lena, right: Fingerprint. Lower row, left: Hill, middle: House, right: Boat.

zero. Observing these distributions shows how important it is to deal with the problem of dimensionality in patch-based image denoising methods.

5. Image filtering by mode estimation

Nowadays, image filtering can be found in a wide range of applications, from multimedia to entertainment and professional imaging. The aim of a good denoising algorithm is to recover the true noiseless image from the noisy observed one. Let us define an image as a set of pixels located on a regular grid. Each pixel of the image is associated with a set of coordinates I. The coordinate vector $\mathbf{i} \in I$ identifies uniquely each pixel. This pixel is called abusively the **i**th pixel. The observed intensity of each pixel is given by

$$x_{\mathbf{i}} = y_{\mathbf{i}} + \varepsilon_{\mathbf{i}}, \tag{4}$$

where $y_{\mathbf{i}}$ is the noisefree pixel and $\varepsilon_{\mathbf{i}}$ is the noise, independent and identically distributed for each pixel. The noise is assumed to be Gaussian with zero mean and standard deviation $v : \varepsilon_{\mathbf{i}} \sim G(0, v^2)$. In practice, a combination of several physical phenomena produces the noise, and the assumptions of independence and normality are often invalidated. However, in most cases, using a Gaussian noise hypothesis provides a simple and powerful background to develop efficient filtering algorithms. Furthermore, if the noise is clearly different from the Gaussian hypothesis, it is often possible to come back to a Gaussian noise problem by using an appropriate variance stabilizing transform (VST) [42], like the Fisz or Anscombe transforms. The aim of an image filtering method is to find $\hat{y}_{\mathbf{i}}$, that is the best possible approximation of $y_{\mathbf{i}}$, from the observed $x_{\mathbf{i}}$, $\forall i \in I$. A good filtering algorithm should have a strong denoising effect while preserving the edges and salient features in the images.

In practice, patch-based filters based on mode estimation like the non-local means [12] can be derived from robust statistics. These filters work by averaging pixels sharing similar neighbourhoods. To identify similar neighbourhoods, similarities between patches are computed using a kernel whose argument is the Euclidean distance between two vectorized image patches. Let us define an image Ω. The distance between the **i**th and **j**th pixel is then defined as $\|\mathbf{i}-\mathbf{j}\|_\infty$. A (square) neighbourhood around the **i**th pixel is defined as $P_{\mathbf{i}} = \{\mathbf{j} \text{ s.t.} \|\mathbf{i}-\mathbf{j}\|_\infty \le r\}$, where r is the radius of the neighbourhood. The intensity of the **i**th pixel is denoted by $x_{\mathbf{i}}$. A patch can then be denoted by $\mathbf{x}_{\mathbf{i}} = [x_{\mathbf{j}}]_{\mathbf{j} \in P_{\mathbf{i}}}$.

Let us write a local kernel density estimation on the patch space $\hat{p}(\mathbf{x}_{\mathbf{i}}) = \sum_{\mathbf{j} \in I} \Psi_\sigma(\|\mathbf{x}_{\mathbf{i}}-\mathbf{x}_{\mathbf{j}}\|_2^2/2)$. The fixed-point update is found by equating to 0 the partial derivative with respect to $x_{\mathbf{k}}$:

$$x_{\mathbf{k}}^{(t+1)} = \frac{\sum_{\mathbf{i} \in P_{\mathbf{k}}} \sum_{\mathbf{j} \in I} w_{\mathbf{ij}} \Psi'_\sigma(\|\mathbf{x}_{\mathbf{i}}^{(t)}-\mathbf{x}_{\mathbf{j}}\|_2^2) x_{\mathbf{j}+\mathbf{i}-\mathbf{k}}^{(t)}}{\sum_{\mathbf{i} \in P_{\mathbf{k}}} \sum_{\mathbf{j} \in I} w_{\mathbf{ij}} \Psi_\sigma{}'(\|\mathbf{x}_{\mathbf{i}}^{(t)}-\mathbf{x}_{\mathbf{j}}\|_2^2)}, \tag{5}$$

where t is the iteration index and $x_{\mathbf{k}}^{(0)}$ is initialized to $x_{\mathbf{k}}$. In images, it is convenient to assume that the content of the images is similar locally only, so most of the important pixels are located in the neighbourhood of the pixel to be filtered. For this reason, $w_{\mathbf{ij}}$, a decaying function of $\|\mathbf{i}-\mathbf{j}\|_2$ chosen in order to keep an acceptable computational load is introduced. This decaying function is usually chosen as a Gaussian kernel with width ρ.

Eq. (5) is in fact an iterative generalization of the well known non-local means [12]. Many variations of this filter exists such as UINTA [14], SAFIR [40], and many others [15–17,41,43,44]. However, most of these variations still use the Gaussian kernel as a measure of similarity, even when using high-dimensional vectors as patches. Some attempts at reducing the effect of the curse of dimensionality can however be found in the literature: [45] uses a principal component analysis in the patch space in order to reduce the dimension. In [15], it is suggested to shift the distribution of patch distances toward zero, just by subtracting a dimensionality-dependent constant from each distance.

In this paper, we propose to fight the curse of dimensionality by replacing the classical Gaussian kernel $\Psi_\sigma(u) = \exp(-u^2/2\sigma^2)$ with

$$\Psi_\sigma(u) = \int_0^u Q\left(\frac{v^2}{4\sigma^2}, \frac{D}{2}\right) v \, dv \tag{6}$$

or by its corresponding Burr type XII approximation. Eq. (6) is heavy to compute and makes the filtering process slower if evaluated for each patch to patch comparison. However, our implementation tabulates the kernel values for a given set of distances before the image is filtered. During the filtering process, when an evaluation of the kernel is needed, the closest precomputed value of the kernel is used. With this method, filtering with the different kernels is virtually equivalent in terms of computational load.

Table 1
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the Lena image.

Lena	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{\text{noise}} = 20$					
Chi	6.3155	0.1035	32.1227	1.3423	7
Gauss	6.7308	0.1057	31.5695	0.7679	7
Burr	6.2168	0.1104	32.2594	1.5843	9
$\sigma_{\text{noise}} = 30$					
Chi	8.8994	0.1854	30.4949	2.0288	9
Gauss	10.1121	0.2443	29.5025	1.1297	7
Burr	9.0007	0.2493	30.4724	2.2849	9
$\sigma_{\text{noise}} = 40$					
Chi	15.8218	0.751	29.1436	2.685	9
Gauss	18.2941	0.7446	28.034	1.529	7
Burr	16.5301	0.7921	29.0453	2.9998	9

6. Experiments

The experiments feature six images with 512×512 pixels and 256 gray levels. These images (Lena, Couple, Hill, House, Fingerprint and Boat images) are widely used in the image processing community[1] (see Fig. 4).

These images are polluted by additive Gaussian white noise with standard deviation $\sigma_{\text{noise}} = 20$, 30 and 40. For each image, different parameters of the filter are tested: k is the patch size varies from 3×3 to 9×9 patches and ρ is fixed as a Gaussian kernel with $\sigma_{w_{\mathbf{ij}}} = 10$ pixels. The width σ of kernel Ψ is optimized in order to minimise the mean RMSE on 100 independent repetitions of the polluted images, for each patch size. The root mean square error (RMSE), its variance and the peak signal to noise ratio (PSNR) are then evaluated on a new set of 100 images, polluted with the same noise model. This procedure is applied on each image, for the traditional Gaussian kernel, the χ CCDF, and the Burr type XII CCDF. Results are reported in Tables 1–6: for each image it presents the mean of the RMSE, its variance and the mean PSNR over the 100 repetitions for the optimal σ and k used to obtain these results. A statistical test has been performed to evaluate the significance of these results: for all images and all noise levels, the denoising effects resulting from the use of the χ, and of the Burr kernels have been tested as significantly better

[1] These images can be downloaded at http://www.cs.tut.fi/~foi/GCF-BM3D/index.html#ref_software.

Table 2
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the Couple image.

Couple	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{noise}=20$					
Chi	8.5052	0.0006	29.5371	1.2306	5
Gauss	8.9471	0.0005	29.0972	0.7488	5
Burr	8.4413	0.0006	29.6026	1.5394	7
$\sigma_{noise}=30$					
Chi	10.4926	0.0012	27.7131	1.9127	7
Gauss	11.4348	0.0011	26.9662	1.1082	5
Burr	10.5253	0.0018	27.686	2.2674	9
$\sigma_{noise}=40$					
Chi	12.1315	0.0017	26.4525	2.5988	9
Gauss	13.4349	0.0011	25.5661	1.5003	5
Burr	12.2984	0.0022	26.3338	2.9443	9

Table 3
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the Fingerprint image.

Fingerprint	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{noise}=20$					
Chi	10.9277	0.0006	27.3602	1.3116	5
Gauss	11.5619	0.0007	26.8702	0.824	5
Burr	10.6613	0.0008	27.5745	1.6983	9
$\sigma_{noise}=30$					
Chi	13.5576	0.0008	25.4871	1.9951	7
Gauss	15.0958	0.0012	24.5537	1.1413	5
Burr	13.392	0.0011	25.5939	2.3645	9
$\sigma_{noise}=40$					
Chi	15.6762	0.0019	24.226	2.6941	9
Gauss	18.3237	0.0027	22.8705	1.4124	9
Burr	15.7793	0.0026	24.169	3.0417	9

Table 4
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the Hill image.

Hill	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{noise}=20$					
Chi	8.1808	0.0004	29.8749	1.2259	5
Gauss	8.5442	0.0004	29.4973	0.7412	5
Burr	8.2256	0.0006	29.8274	1.5051	7
$\sigma_{noise}=30$					
Chi	9.8072	0.0007	28.2998	1.9584	9
Gauss	10.4783	0.0007	27.725	1.128	5
Burr	9.8617	0.0009	28.2518	2.2206	9
$\sigma_{noise}=40$					
Chi	15.8218	0.2138	24.142	2.4776	9
Gauss	18.2941	0.1625	22.8825	1.6768	5
Burr	16.5301	0.1749	23.7626	3.1002	9

than in the Gaussian kernel case ($p<0.01$ on all cases). The optimal value of σ is larger in the case of the flat-top kernels, because of the steeper slope of these kernels. The difference between the results increases as the noise increases, suggesting that the gain of performances is more important as the noise is intense.

Table 5
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the House image.

House	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{noise}=20$					
Chi	6.0352	0.0026	32.5167	1.3148	5
Gauss	6.4571	0.0026	31.9297	0.8258	5
Burr	5.9574	0.0033	32.6293	1.5726	9
$\sigma_{noise}=30$					
Chi	7.42	0.0064	30.7222	2.026	9
Gauss	8.5187	0.0057	29.523	1.1927	5
Burr	7.4609	0.0064	30.6744	2.2742	9
$\sigma_{noise}=40$					
Chi	8.8442	0.0103	29.197	2.6423	9
Gauss	10.4509	0.0091	27.7473	1.6001	5
Burr	9.0149	0.0103	29.031	2.9645	9

Table 6
Mean root mean square error, variance of root mean square error, mean peak signal to noise ratio, optimal σ, and optimal patch size for 100 repetitions of the Boat image.

Boat	*RMSE*	var(*RMSE*)	*PSNR*	σ	k
$\sigma_{noise}=20$					
Chi	8.3176	0.0005	29.7308	1.2466	5
Gauss	8.6265	0.0004	29.4141	0.7645	5
Burr	8.268	0.0004	29.7828	1.5143	5
$\sigma_{noise}=30$					
Chi	10.1054	0.0011	28.0397	1.933	7
Gauss	10.8238	0.0008	27.4432	1.1342	5
Burr	10.1124	0.0013	28.0337	2.2782	9
$\sigma_{noise}=40$					
Chi	11.6133	0.0029	26.8316	2.635	9
Gauss	12.7309	0.0019	26.0336	1.5317	5
Burr	11.6831	0.0026	26.7796	2.9844	9

Example results for the Lena image are shown in Figs. 5–7. The images were cropped in order to be able to better see the fine details of the images. For all noise levels, the visual quality is slightly better when the flat-top kernels are used instead of the Gaussian kernel. This impression gets stronger as the noise level increases, as the RMSE and PSNR suggest.

7. Conclusions

Mode estimation is a useful tool in signal processing and image denoising, especially for denoising tasks. Modes are typically searched by running a hill-climbing procedure on the kernel density estimator. This iterative approach can be used in mono-dimensional as well as multidimensional spaces. However, this paper has shown that the generalization to several dimensions is not as straightforward as it seems at first glance. More specifically, the widely used Gaussian kernel, interpreted as a local similarity measure, no longer appears to be optimal and shows a poor discrimination power. Taking into account the phenomenon of norm concentration suggests on the other hand that flat-top kernels are more appropriate in high-dimensional spaces.

The theoretical developments are supported and illustrated with an image denoising application. Data are generated by decomposing pictures into overlapping patches, which are then vectorized. Denoising is performed by replacing each patch with

Fig. 5. Examples of results obtained for the Lena image with $\sigma_{\text{noise}} = 20$. Upper row, left: noisy image. Right: denoised image with Gaussian kernel. Lower row, left: denoised image with Chi kernel. Right: denoised image with Burr kernel.

Fig. 6. Examples of results obtained for the Lena image with $\sigma_{\text{noise}} = 30$. Upper row, left: noisy image. Right: denoised image with Gaussian kernel. Lower row, left: denoised image with Chi kernel. Right: denoised image with Burr kernel.

Fig. 7. Examples of results obtained for the Lena image with $\sigma_{\text{noise}} = 40$. Upper row, left: noisy image. Right: denoised image with Gaussian kernel. Lower row, left: denoised image with Chi kernel. Right: denoised image with Burr kernel.

its closest mode in the observed image-specific patch distribution. Eventually, the resulting patches are reassembled into a denoised image. The experimental section shows that the kernel shape has a significant impact on quantitative performance measures. The experiments feature six different images, three noise levels, and 100 noise samples for three different kernels. In each case, the flat-top kernels outperform the Gaussian kernel.

References

[1] E. Parzen, On estimation of a probability density function and mode, The Annals of Mathematical Statistics 33 (3) (1962) 1065–1076.

[2] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8) (1995) 790–799.

[3] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.

[4] R. van den Boomgaard, J. van de Weijer, On the equivalence of local-mode finding, robust estimation and mean-shift analysis as used in early vision tasks, in: IEEE Conference on Pattern Recognition, vol. 3, Quebec, Canada, 2002, pp. 927–930.

[5] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics, Wiley Series in Probability and Mathematical Statistics, Wiley & Sons, New York, 1986.

[6] P.J. Huber, Robust Statistics, Wiley Series in Probability and Mathematical Statistics, Wiley & Sons, New York, 1981.

[7] C. Chu, I. Glad, F. Godtliebsen, M. J.S., Edge-preserving smoothers for image processing, Journal of the American Statistical Association 93 (442) (1998) 526–556.

[8] G. Winkler, V. Aurich, K. Hahn, A. Martin, K. Rodenacker, Noise reduction in images: some recent edge-preserving methods, Mathematik, Informatik und Statistik, Statistik, Sonderforschungsbereich, 1998.

[9] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, in: International Conference on Computer Vision, Bombay, India, 1998, pp. 893–846.

[10] M. Elad, On the origin of the bilateral filter and ways to improve it, IEEE Transactions on Image Processing 11 (10) (2002) 1141–1151.

[11] D. Scott, Multivariate Density Estimation: Theory, Practice and Visualization, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, 1992.

[12] A. Buades, B. Coll, J. Morel, A review of image denoising algorithms with a new one, Multiscale Modeling & Simulation 4 (2) (2005) 490–530.

[13] A. Buades, B. Coll, J.-M. Morel, The staircasing effect in neighborhood filters and its solution, IEEE Transactions on Image Processing 15 (6) (2006) 1499–1505.

[14] S.A. Awate, R.T. Whitaker, Image denoising with unsupervised, information-theoric, adaptative filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (3) (2006) 364–376.

[15] C. Kervrann, J. Boulanger, P. Coupé, Bayesian non-local means filter, image redundancy and adaptative dictionaries for noise removal, in: Conference on Scale-Space and Variational Methods, vol. 4485, Ischia, Italy, 2007, pp. 520–532.

[16] C. Kervrann, J. Boulanger, Optimal spatial adaptation for patch-based image denoising, IEEE Transactions on Image Processing 15 (10) (2006) 2866–2878.

[17] C.-A. Deledalle, L. Denis, F. Tupin, Iterative weighted maximum likelihood denoising with probabilistic patch-based weights, IEEE Transactions on Image Processing 18 (12) (2009) 2661–2672.

[18] B.U. Park, B.A. Turlach, Practical performance of several data driven bandwidth selectors, Computational Statistics 7 (1992) 251–270.

[19] R. Cao, A. Cuevas, W.G. Manteiga, A comparative study of several smoothing methods in density estimation, Computational Statistics and Data Analysis 17 (1994) 153–176.

[20] M. Farmen, J.S. Marron, An assessment of finite sample performance of adaptive methods in density estimation, Computational Statistics and Data Analysis 30 (1999) 143–168.

[21] R. Bellman, Adaptative Control Processes: A Guided Tour, Princeton University Press, Princeton, NJ, 1961.

[22] D. Donoho, High-dimensional data analysis: the curse and blessings of dimensionality, Aide-memoire for a Lecture for the American Mathematical Society: Mathematical Challenges of the 21st Century, 2000.

[23] D. Francois, V. Wertz, M. Verleysen, The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering 19 (7) (2007) 873–886.

[24] R. Charnigo, J. Sun, R. Muzic, A semi-local paradigm for wavelet denoising, IEEE Transactions on Image Processing 15 (3) (2006) 666–677.

[25] R. Coifman, D. Donoho, Translation invariant denoising, Wavelets and Statistics, 1995, pp. 125–150.

[26] D. Donoho, I. Johnstone, Ideal spatial adaptation via wavelet shrinkage, Biometrika 81 (1994) 425–455.

[27] P. Perona, J. Malik, Scale-space and edge-detection using anisotropic diffusion, Transactions on Pattern Analysis and Machine Intelligence 12 (7) (1990) 629–639.
[28] M. Black, G. Sapiro, D. Marimont, D. Heeger, Robust anisotropic diffusion, IEEE Transactions on Image Processing 7 (3) (1998) 421–432.
[29] J. Weickert, Anisotropic Diffusion in Image Processing, ECMI Series, 1998.
[30] S. Osher, M. Burger, G.D.X.J.W. Yin, An iterative regularization method for total variation-based image restoration, Multiscale Modelling & Simulation 4 (2) (2005) 460–489.
[31] C. Vogel, M. Oman, Iterative methods for total variation denoising, Journal on Scientific Computing 17 (1) (1996) 227–238.
[32] E. Tadmor, S. Nezzar, L. Vese, A multiscale image representation using hierarchical (BV,L^2) decompositions, Multiscale Modeling & Simulation 2 (4) (2004) 554–579.
[33] K. Dabov, A. Foi, K. Egiazarian, Image restoration by sparse 3d transform-domain collaborative filtering, no. 6812-07, San Jose, California, USA, 2008.
[34] A. de Decker, J. Lee, D. Francois, M. Verleysen, Mode estimation in high-dimensional spaces with flat-top kernels: application to image denoising, in: European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 2010, pp. 411–417.
[35] D. Francois, V. Wertz, M. Verleysen, About the locality of kernels in high-dimensional spaces, in: International Symposium on Applied Stochastic Models and Data Analysis, Brest, France, 2005, pp. 238–245.
[36] D. Francois, High-dimensional Data Analysis: From Optimal Metrics to Feature Selection, Verlag-VDM, 2008.
[37] G. Arfken, The Incomplete Gamma Function and Related Functions, third ed., vol. 10.5, 1985.
[38] D. Francois, V. Wertz, M. Verleysen, About the locality of kernels in high-dimensional spaces, in: Proceedings of ASMDA 2005, International Symposium on Applied Stochastic Models and Data Analysis, Brest, France, 17–19 May 2005, pp. 238–245.
[39] I. Burr, Cumulative frequency functions, Annals of Mathematical Statistics 13 (1942) 215–232.
[40] C. Kervrann, J. Boulanger, Unsupervised patch-based image regularization and representation, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Graz, Austria, 2006, pp. 555–567.
[41] B. Goossens, H. Luong, A. Pizurica, W. Philips, An improved non-local denoising algorithm, in: International Workshop on Local and Non-Local Approximation in Image Processing, Lausanne, Switzerland, 2008, pp. 25–29.
[42] J.A. Lee, X. Geets, V. Gregoire, A. Bol, Edge-preserving filtering of images with low photon counts, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6) (2008) 1014–1027.
[43] P. Chatterjee, P. Milanfar, A generalization of non-local means via kernel regression, in: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 6814, 2008.
[44] T. Brox, O. Kleinschmidt, D. Cremers, Efficient non-local means for denoising of textural patterns, IEEE Transactions on Image Processing 17 (7) (2008) 1083–1092.
[45] T. Tasdizen, Principal neighborhood dictionaries for non-local means image denoising, IEEE Transactions on Image Processing 18 (12) (2009) 2649–2660.

Arnaud de Decker was born in Brussels, Belgium, in 1982. He received his M.Sc. degree in physics (space, climate and earth physics) in 2005 and an M.Sc. in statistics in 2006, both from the Université catholique de Louvain (UCL, Belgium).

His main research interests are ECG automatic annotation, medical image denosing and deblurring, and automatic tumor and organs delineation on medical images.

He is now a Ph.D. student in the machine learning group of the Polytechnic School at UCL.

Damien François received the engineering degree in computer science in 2002 and the Ph.D. degree in applied mathematics in 2007 from the Université catholique de Louvain, Belgium. His main research interests are in the field of high-dimensional data analysis, feature selection, and meta learning, mainly applied to biomedical/life sciences applications and business applications.

Michel Verleysen was born in 1965 in Belgium. He received the M.S. and the Ph.D. degrees in Electrical Engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively.

He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (E.P.F.L.), Switzerland in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1—Panthéon-Sorbonne in 2002–2004. He is a former Research Director with the Belgian FNRS (Fonds National de la Recherche Scientifique) and a Professor at the Université catholique de Louvain.

He is Editor-in-Chief of the Neural Processing Letters journal, Chairman of the Annual European Symposium on Artificial Neural Networks (ESANN) Conference, Associate Editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning.

He is the author or the co-author of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee.

He is the co-author of the scientific popularization book on artificial neural networks in the series 'Que Sais-Je?,' in French.

His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.

John Aldo Lee was born in 1976 in Brussels, Belgium. He received the M.Sc. degree in Applied Sciences (Computer Engineering) in 1999 and the Ph.D. degree in Applied Sciences (Machine Learning) in 2003, both from the Université catholique de Louvain (UCL, Belgium).

His main interests are nonlinear dimensionality reduction, intrinsic dimensionality estimation, independent component analysis, clustering, and vector quantization.

He is a member of the UCL Machine Learning Group and is now a Research Associate with the Belgian FNRS (Fonds National de la Recherche Scientifique).

His current work aims at developing specific image enhancement techniques for positron emission tomography in the Molecular Imaging and Experimental Radiotherapy Department of the Saint-Luc University Hospital (Belgium).