Multi image noise estimation and denoising
Antoni Buades, Yifei Lou, Jean-Michel Morel, Zhongwei Tang

To cite this version:
Antoni Buades, Yifei Lou, Jean-Michel Morel, Zhongwei Tang. Multi image noise estimation and denoising. 2010. hal-00510866

HAL Id: hal-00510866
https://hal.archives-ouvertes.fr/hal-00510866
Submitted on 22 Aug 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Multi image noise estimation and denoising
A. Buades∗ Y. Lou† J.M Morel‡ Z. Tang‡
August 22, 2010

Abstract
Photon accumulation on a fixed surface is the essence of photography. In the times of chemical photography this accumulation required the camera to move as little as possible, and the scene to be still. Yet, most recent reflex and compact cameras propose a burst mode, permitting to capture quickly dozens of short exposure images of a scene instead of a single one. This new feature permits in principle to obtain by simple accumulation high quality photographs in dim light, with no motion or aperture blur. It also gives the right data for an accurate noise model. Yet, both goals are attainable only if an accurate cross-registration of the burst images has been performed. The difficulty comes from the non negligible image deformations caused by the slightest camera motion, in front of a 3D scene, and from the light variations or motions in the scene. This paper proposes a numerical processing chain permitting to achieve jointly the two mentioned goals: an accurate noise model for the camera, which is used crucially to obtain a state of the art multi-images denoising. The key feature of the proposed processing chain is a reliable multi-image noise estimator, whose accuracy will be demonstrated by three different procedures. Thanks to the signal dependent noise model obtained from the burst itself, a faithful detection of the well registered pixels can be made. The denoising by simple accumulation of these pixels, which are an overwhelming majority, permits to extend the Nicéphore Niepce photon accumulation method to image bursts. The denoising performance by accumulation is shown to reach the theoretical limit, namely a $\sqrt{n}$ denoising factor for $n$ frames. Comparison with state of the art denoising algorithms will be shown on several bursts taken with reflex cameras in dim light.

1 Introduction
The accumulation of photon impacts on a surface is the essence of photography. The first Nicephore Niepce photograph [20] was obtained after an eight hours exposure. The serious objection to a long exposure is the variation of the scene due to changes in light, camera motion, and incidental motions of parts of the scene. The more these variations can be compensated, the longer the exposure can be, and the more the noise can be reduced. It is a frustrating experience for professional photographers to take pictures under bad lighting conditions with a hand-held camera. If the camera is set to a long exposure time, the photograph gets blurred by the camera motions and aperture. If it is taken with short exposure, the image is dark, and enhancing it reveals the noise. Yet,

∗MAP5, CNRS - Université Paris Descartes, 45 rue Saints Peres, 75270 Paris Cedex 06, France.
†Mathematics Department, University of California at Los Angeles, Los Angeles, U.S.A.
‡CMLA, ENS Cachan, 61 av. President Wilson, Cachan 94235, France.
this dilemma can be solved by taking a burst of images, each with short-exposure time, as shown in Fig. 1, and by averaging them after registration. This observation is not new and many algorithms have been proposed, mostly for stitching and super-resolution. These algorithms have thrived in the last decade, probably thanks to the discovery of a reliable algorithm for image matching, the SIFT algorithm [55]. All of the multi-image fusion algorithms share three well separated stages, the search and matching of characteristic points, the registration of consecutive image pairs and the final accumulation of images. All methods perform some sort of multi-image registration, but surprisingly do not propose a procedure to check if the registration is coherent. Thus, there is a non-controlled risk that the accumulation blurs the final accumulation image, due to wrong registrations. Nevertheless, as we shall see, the accurate knowledge of noise statistics for the image sequence permits to detect and correct all registration incoherences. Furthermore, this noise statistics can be most reliably extracted from the burst itself, be it for raw or for JPEG images. In consequence, a stand alone algorithm which denoises any image burst is doable. As experiments will show, it even allows for light variations and moving objects in the scene, and it reaches the $\sqrt{n}$ denoising factor predicted for the sum of the $n$ independent (noise) random variables.

We call in the following “burst”, or “image burst” a set of digital images taken from the same camera, in the same state, and quasi instantaneously. Such bursts are obtained by video, or by using the burst mode proposed in recent reflex and compact cameras. The camera is supposed to be held as steady as possible so that a large majority of pixels are seen through the whole burst. Thus, no erratic or rash motion of the camera is allowed, but instead incident motions in the scene do not hamper the method.

There are other new and promising approaches, where taking images with different capture conditions is taken advantage of. Liu et al. [88] combine a blurred image with long-exposure time, and a noisy one with short-exposure time for the purpose of denoising the second and deblurring the first. Beltramo and Levine [11] improve the dynamic range of the final image by combining an underexposed snapshot with an overexposed one. Combining again two snapshots, one with and the other without flash, is investigated by Eisemann et. al. [33] and Fattal et. al [37]. Another case of image fusion worth mentioning is [8], designed for a 3D scanning system. During each photography session, a high-resolution digital back is used for photography, and separate macro (close-up) and ultraviolet light shots are taken of specific areas of text. As a result, a number of folios are captured with two sets of data: a “dirty” image with registered 3D geometry and a “clean” image with the page potentially deformed differently to which the digital flattening algorithms are applied.

Our purpose here is narrower. We only aim at an accurate noise estimation followed by denoising for an image burst. No super-resolution will be attempted, nor the combination of images taken under different apertures, lightings or positions. The main assumption on the setting is that a hand-held camera has taken an image burst of a still scene, or from a scene with a minority of moving objects. To get a significant denoising, the number of images can range from 9 to 64, which grants a noise reduction by a factor 3 to 8. Since the denoising performance grows like the square root of the number of images, it is less and less advantageous to accumulate images when their number grows. But impressive denoising factors up to 6 or 8 are reachable by the simple algorithm proposed here, which we shall call average after registration (AAR). Probably the closest precursor to the present method is the multiple image denoising method by Zhang et. al. [92]. Their images are not the result of a burst. They are images taken from different points of views by different cameras. Each camera uses a small aperture and a short exposure to ensure
Plan and results of the paper. The paper requires a rich bibliographical analysis for the many aspects of multi-image processing (Section 3). This survey shows that most super-resolution algorithms do in fact much more denoising than they do super-resolution, since they typically only increase the size of the image by a factor 2 or 3, while the number of images would theoretically allow for a 5 to 8 factor. Section 2 reviews the other pillar of the proposed method, the noise estimation literature. (This corpus is surprisingly poor in comparison to the denoising literature.)

Section 4 is key to the proposed technique, as it demonstrates that a new variant of static noise blind estimate gives results that exactly coincide with Poisson noise estimates taken from registered images in a temporal sequence. It is also shown that although JPEG images obtained by off-the-shelf cameras have no noise model, a usable substitute to this noise model can be obtained: It simply is the variance of temporal sequences of registered images. Section 5 describes the proposed multi-image denoising method, which in some sense trivializes the denoising technology, since it proposes to go back as much as possible to a mere accumulation, and to perform a more sophisticated denoising only at dubiously registered pixels. Section 6 compares the proposed strategy with two state of the art multi-images denoising strategies.

2 Noise Estimation, a review

As pointed out in [53], “Compared to the in-depth and wide literature on image denoising, the literature on noise estimation is very limited”. Following the classical study by Healey et al. [45], the noise in CCD sensors can be approximated by an additive, white and signal dependent noise model. The noise model and its variance reflect different aspects of the imaging chain at the CCD, mainly dark noise and shot noise. Dark noise is due to the creation of spurious electrons generated by thermal energy which become indistinguishable from photoelectrons. Shot noise is a result of the quantum nature of light and characterizes the uncertainty in the number of photons stored at a collection site. This number of photons follows a Poisson distribution so that its variance equals its mean. The overall combination of the different noise sources therefore leads to an affine
Figure 1: From left to right: one long-exposure image (time = 0.4 sec, ISO=100), one of 16 short-exposure images (time = 1/40 sec, ISO = 1600) and the average after registration. All images have been color-balanced to show the same contrast. The long exposure image is blurry due to camera motion. The middle short-exposure image is noisy, and the third one is some four times less noisy, being the result of averaging 16 short-exposure images. Images may need to be zoomed in on a screen to compare details and textures.

noise variance $a + bu$ depending on the original signal value $u$. Yet, this is only true for the raw CCD image. Further processing stages in the camera hardware and software such as the white balance, the demosaicking, the gamma correction, the blur and color corrections, and eventually the compression, correlate the noise and modify its nature and its standard deviation in a non trivial manner. There is therefore no noise model for JPEG images. However, as we shall see, a signal dependent noise variance model can still be estimated from bursts of JPEG images (section 4.2.) It is enough to perform reliably the average after registration (AAR).

2.1 Additive Gaussian noise estimation

Most computer vision algorithms should adjust their parameters according to the image noise level. Surprisingly, there are few papers dealing with the noise estimation problem, and most of them only estimate the variance of a signal independent additive white Gaussian noise (AWGN). This noise statistics is typically measured on the highest-frequency portion of the image spectrum, or on homogenous image patches. In the AWGN case a spectral decomposition through an orthonormal transform such as wavelets or the DCT preserves the noise statistics. To estimate the noise variance, Donoho et. al [29] consider the finest scale wavelet coefficients, followed by a median filter to eliminate the outliers. Suppose \( \{y_i\}_{i=1,...,N} \) be \( N \) independent Gaussian random variables of zero-mean and variance \( \sigma^2 \), then

\[
E\{\text{MED}(|y_i|)\} \approx 0.6745\sigma.
\]

It follows immediately that the noise standard deviation \( \sigma \) is given by

\[
\hat{\sigma} = \frac{1}{0.6745} \text{MED}(|y_i|) = 1.4826 \text{MED}(|y_i|).
\]

The standard procedure of the local approaches is to analyze a set of local estimates of the variance. For example, Rank et. al [48] take the maximum of the distribution of image derivatives. This method is based on the assumption that the underlying image has a large portion of homogeneous regions. Yet, if an image is highly textured, the noise
variance can be overestimated. To overcome this problem, Ponomarenko et al [72] have proposed to analyze the local DCT coefficients. A segmentation-based noise estimation is carried out in [1], which considers both i.i.d. and spatially correlated noise.

The algorithm in [73] is a modification of the early work [72] dealing with AVIRIS (Airborne Visible Infrared Imaging Spectrometer) images, in which the evaluation of the noise variance in sub-band images is addressed. The idea is to divide each block into low frequency and high frequency components by thresholding, and to use $K$ blocks of the smallest variance of the low frequency coefficients to calculate a noise variance, where $K$ is adaptively selected so that it is smaller for highly-textured images.

[25] proposed an improvement of the estimate of the variance of AWGN by using transforms creating a sparse representation (via BM3D [22]) and using robust statistics estimators (MAD and ICI). For a univariate data set $X_1, X_2, ..., X_n$, the MAD is defined as the median of the absolute deviations from the data’s median: $\text{MAD} = \text{median}_i \left( |X_i - \text{median}_j(X_j)| \right)$. The algorithm is as follows.

1. for each $8 \times 8$ block, group together up to 16 similar non-overlapping blocks into 3D array. The similarity between blocks in evaluated by comparing corresponding blocks extracted from a denoised version by BM3D.
2. apply a 3-D orthonormal transform (DCT or wavelet) on each group and sort the coefficients according to the zig-zag scan.
3. collect the first 6 coefficients $c_1, \cdots, c_6$ and define their empirical energy as the mean of the magnitude of the (up to 32) subsequent coefficients:
   $$E(|c_j|^2) = \text{mean}\{ |c_{j+1}^2|, \cdots, c_{j+32}^2| \}$$
4. Sort the coefficients from all the groups (6 coefficients per group) according to their energy
5. do MAD and Intersection of Confidence Intervals (ICI) [42] to achieve the optimal bias-variance trade-off in the MAD estimation.

All the above mentioned algorithms give reasonable estimates of the standard deviation when the noise is uniform. Yet, when applying these algorithms to estimate signal dependent noise, the results are poor. The work of C. Liu et. al. [54] estimates the upper bound on the noise level fitting to a camera model. The noise estimation from the raw data is discussed in [39, 40]. The former is a parametric estimation by fitting the model to the additive Poissonian-Gaussian noise from a single image, while the latter measures the temporal noise based on an automatic segmentation of 50 images.

### 2.2 Poisson Noise Removal

This paper deals with real noise, which in most real images (digital cameras, tomography, microscopy and astronomy) is a Poisson noise. The Poisson noise is inherent to photon counting. This noise adds up to a thermal noise and an electronic noise which are approximately AWGN. In the literature algorithms considering the removal of AWGN are dominant but, if its model is known, Poisson noise can be approximately reduced to AWGN by a so called variance stabilizing transformation (VST). The standard procedure follows three steps,

1. apply VST to make the data homoscedastic
2. denoise the transformed data
3. apply the inverse VST.

The square-root operation is widely used as a VST,

\[ f(z) = b\sqrt{z} + c. \]  

(1)

It follows from the asymptotic unit variance of \( f(z) \) that the parameters are given by \( b = 2 \) and \( c = 3/8 \), which is the Anscombe transform [2]. A multiscale VST (MS-VST) is studied in [91] along with the conventional denoising schemes based on wavelets, ridgelets and curvelets depending on morphological features (isotropic, line-like, curvilinear, etc) of the given data. It is argued in [58] that the inverse transformation of VST is crucial to the denoising performance. Both the algebraic inverse

\[ I_A(D) = \left( \frac{D}{2} \right)^2 - \frac{3}{8}, \]

and the asymptotically unbiased inverse

\[ I_B(D) = \left( \frac{D}{2} \right)^2 - \frac{1}{8}, \]

in [2] are biased for low counts. The authors [58] propose an exact unbiased inverse. They consider an inverse transform \( I_C \) that maps the value \( E\{f(z)|y\} \) to the desired value \( Ez|y \) that

\[ E\{f(z)|y\} = 2 \sum_{z=0}^{\infty} \left( \sqrt{z + \frac{3}{8}} \frac{y^z \exp^{-y}}{z!} \right) \]

where \( f(z) \) is the forward Anscombe transform (1). In practice, it is sufficient to compute the above equation for a limited set of values \( y \) and approximate \( I_C \) by \( I_B \) for large values of \( y \). Furthermore, the state-of-the-art denoising scheme BM3D [39] is applied in the second step.

There are also wavelets based methods [69, 50] or Bayesian [80, 56, 51] removing Poisson noise. In particular, the wavelet-domain Wiener filter [69] uses a cross-validation that not only preserves important image features, but also adapts to the local noise level of the spatially varying Poisson process. The shrinkage of wavelet coefficients investigates how to correct the thresholds [50] to explicitly account for effects of the Poisson distribution on the tails of the coefficient distributions. A recent Bayesian approach by Lefkimmiatis et al. [51] explores a recursive quad-tree image representation which is suitable for Poisson noise degradation and then follows an expectation-maximization technique for parameter estimation and Hidden Markov tree (HMT) structures for inter-scale dependencies. The common denominator to all such methods is that we need an accurate Poisson model, and this will be thoroughly discussed in Section 4.

It is, however, a fact that the immense majority of accessible images are JPEG images which contain a noise altered by a long chain of processing algorithms, ending with compression. Thus the problem of estimating noise in a single JPEG image is extremely ill-posed. It has been the object of a thorough study in [53]. This paper proposes a blind estimation and removal method of color noise from a single image. The interesting feature is that it constructs a “noise level function” which is signal dependent, obtained by computing empirical standard deviations image homogeneous segments. Of course the remanent noise in a JPEG image is no way white or homogeneous, the high frequencies
being notoriously removed by the JPEG algorithm. On the other hand, demosaicking usually acts as a villainous converter of white noise into very structured colored noise, with very large spots. Thus, even the variance of smooth regions cannot give a complete account of the noise damage, because noise in JPEG images is converted in extended flat spots. We shall, however, retain the idea promoted in [53] that, in JPEG images, a signal dependent model for the noise variance can be found. In section 4.2 a simple algorithm will be proposed to estimate the color dependent variance in JPEG images from multi-images. All in all, the problem of estimating a noise variance is indeed much better posed if several images of the same scene by the same camera, with the same camera parameters, are available. This technique is classic in lab camera calibration [44].

3 Multi-images and super resolution algorithms

Photo stitching  Probably one of the most popular applications in image processing, photo stitching [14, 57] is the first method to have popularized the SIFT method permitting to register into a panorama a set of image of a same scene. Another related application is video stabilization [7]. In these applications no increase in resolution is gained, the final image has roughly the same resolution as the initial ones.

Super-resolution  Super-resolution means creating a higher resolution, larger image from several images of the same scene. Thus, this theme is directly related to the denoising of image bursts. It is actually far more ambitious, since it involves a deconvolution. However, we shall see that most super-resolution algorithms actually make a moderate zoom in, out of many images, and therefore mainly perform a denoising by some sort of accumulation. The convolution model in the found references is anyway not accurate enough to permit a strong deconvolution.

A single-frame super-resolution is often referred to as interpolation. See for example [85, 86]. But several exemplar-based super-resolution methods involve other images which are used for learning, like in Baker and Kanade [4] who use face or text images as priors. Similarly, the patch-example-based approaches stemming from the seminal paper [41], use a nearest-neighbor search to find the best match for local patches, and replace them with the corresponding high-resolution patches in the training set, thus enhancing the resolution. To make the neighbors compatible, the belief-propagation algorithm to the Markov network is applied, while another paper [26] considered a weighted average by surrounding pixels (analogue to nonlocal means [15]). Instead of a nearest-neighbor search, Yang et. al [83] proposed to incorporate the sparsity in the sense that each local patch can be sparsely represented as a linear combination of low-resolution image patches; and a high-resolution image is reconstructed by the corresponding high-resolution elements. The recent remarkable results of [87] go in the same direction. The example-based video enhancement is discussed in [12], where a simple frame-by-frame approach is combined with temporal consistency between successive frames. Also to mitigate the flicker artifacts, a stasis prior is introduced to ensure the consistency in the high frequency information between two adjacent frames.

Focus on registration  In terms of image registration, most of the existing super-resolution methods rely either on a computationally intensive optical flow calculation, or on a parametric global motion estimation. The authors of [94] discuss the effects of multi-image alignment on super-resolution. The flow algorithm they employ addresses
two issues: flow consistency (flow computed from frame A to frame B should be consistent with that computed from B to A) and flow accuracy. The flow consistency can be generalized to many frames by computing a consistent bundle of flow fields. Local motion is usually estimated by optical flow, other local deformation models include Delaunay triangulation of features [9] and B-splines [64]. Global motion, on the other hand, can be estimated either in the frequency domain or by feature-based approaches. For example, Vandewalle et. al. [82] proposed to register a set of images based on their low-frequencies, aliasing-free part. They assume a planar motion, and as a result, the rotation angle and shifts between any two images can be precisely calculated in the frequency domain. The standard procedure for feature-based approaches is (1) to detect the key points via Harris corner [19, 3] or SIFT [89, 75], (2) match the corresponding points while eliminating outliers by RANSAC and (3) fit a proper transformation such as a homography. The other applications of SIFT registration are listed in Tab. 2.

Reconstruction after registration A number of papers focus on image fusion, assuming the motion between two frames is either known or easily computed. Elad and Feuer [34] formulate the super-resolution of image sequences in the context of Kalman filtering. They assume that the matrices which define the state-space system are known. For example, the blurring kernel can be estimated by a knowledge of the camera characteristics, and the warping between two consecutive frames is computed by a motion estimation algorithm. But due to the curse of dimensionality of the Kalman filter, they can only deal with small images, e.g., of size $50 \times 50$. The work [59] by Marquina and Osher limited the forward model to be spatial-invariant blurring kernel with the down-sampling operator, while no local motion was present. They solved a TV-based reconstruction with Bregman iterations.

A joint approach on demosaicing and super-resolution of color images is addressed in [35], based on their early super-resolution work [36]. The authors use the bilateral-TV regularization for the spatial luminance component, the Tikhonov regularization for the chrominance component and a penalty term for inter-color dependencies. The motion vectors are computed via a hierarchical model-based estimation [10]. The initial guess is the result of the Shift-And-Add method. In addition, the camera PSF is assumed to be a Gaussian kernel with various standard deviation for different sets of experiments.

Methods joining denoising, deblurring, and motion compensation Super-resolution and motion deblurring are crossed in the work [5]. First the object is tracked through the sequence, which gives a reliable and sub-pixel segmentation of a moving object [6]. Then a high-resolution is constructed by merging the multiple images with the motion estimation. The deblurring algorithm, which mainly deals with motion blur [47], has been applied only to the region of interest. The recent paper on super-resolution by L. Baboulaz and P. L. Dragotti [3] presents several registration and fusion methods. The registration can be performed either globally by continuous moments from samples, or locally by step edge extraction. The set of registered images is merged into a single image to which either a Wiener or an iterative Modified Residual Norm Steepest Descent (MRNSD) method is applied [67] to remove the blur and the noise. The super-resolution in [75] uses SIFT + RANSAC to compute the homography between the template image and the others in the video sequence, shifts the low-resolution image with subpixel accuracy and selects the closest image with the optimal shifts.
Implicit motion estimation  More recently, inspired by the nonlocal movie denoising method, which claims that “denoising images sequences does not require motion estimation” [16], researchers have turned their attention towards super-resolution without motion estimation [32, 31, 74]. Similar methodologies include the steering kernel regression [78], BM3D [24] and its many variants. The forward model in [24] does not assume the presence of the noise. Thus the authors pre-filter the noisy LR input by V-BM3D [21]. They up-sample each image progressively m times, and at each time, the initial estimate is obtained by zero-padding the spectra of the output from the previous stage, followed by filtering. The overall enlargement is three times the original size. Super-resolution in both space and time is discussed in [76, 77], which combine multiple low-resolution video sequences of the same dynamic scene. They register any two sequences by a spatial homography and a temporal affine transformation, followed by a regularization-based reconstruction algorithm.

A synoptic table of super-resolution multi-images methods  Because the literature is so rich, a table of the mentioned methods, classified by their main features, is worth looking at. The methods can be characterized by a) their number $k$ of fused images, which goes from 1 to 60, b) the zoom factor, usually 2 or 3, and therefore far inferior to the potential zoom factor $\sqrt{k}$, c) the registration method, d) the deblurring method, e) the blur kernel. A survey of the table demonstrates that a majority of the methods use many images to get a moderate zoom, meaning that the denoising factor is important. Thus, these methods denoise in some sense by accumulation. But, precisely because all of them aim at super-resolution, none of them considers the accumulation by itself.

Tables 1 and 2 confirm the dominance of SIFT+RANSAC as a standard way to register multi-images, as will also be proposed here in an improved variant. Several of the methods in Table 1 which do not perform SIFT+RANSAC, actually the last four rows, are “implicit”. This means that they adhere to the dogma that denoising does not require motion estimation. It is replaced by multiple block motion estimation, like the one performed in NL-means and BM3D. However, we shall see in the experimental section that AAR (average after registration) has a still better performance than such implicit methods. This is one of the main questions that arose in this exploration, and the answer is clear cut: denoising by accumulation, like in ancient photography times still is a valid response in the digital era.

4 Noise blind estimation

In this section we return to noise estimation and will confront and cross-validate a single frame noise estimation with a multi-images noise estimation.

4.1 Single image noise estimation

Most noise estimation methods have in common that the noise standard deviation is computed by measuring the derivative or equivalently the wavelet coefficient values of the image. As we mentioned, Donoho et al. [30] proposed to estimate the noise standard deviation as the median of absolute values of wavelet coefficients at the finest scale. Instead of the median, many authors [13, 49] prefer to use a robust median.

Olsen [70] and posteriorly Rank et al. [48] proposed to compute the noise standard deviation by taking a robust estimate on the histogram of sample variances of patches
Table 1: comparison of Super Resolution algorithms

<table>
<thead>
<tr>
<th>Ref.</th>
<th># of images</th>
<th>Registration</th>
<th>Deblurring</th>
<th>blur kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>[41]</td>
<td>1</td>
<td>2</td>
<td>KNN to training set</td>
<td>NO</td>
</tr>
<tr>
<td>[4]</td>
<td>to 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[26]</td>
<td>1</td>
<td>2</td>
<td>MAP penalty</td>
<td>3 × 3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>5 × 5</td>
</tr>
<tr>
<td>[83]</td>
<td>1</td>
<td>to 4</td>
<td>sparse w.r.t. traning</td>
<td>back-projection</td>
</tr>
<tr>
<td>[19]</td>
<td>15</td>
<td>2</td>
<td>Harris+RANSAC</td>
<td>Tiknonov</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[18]</td>
<td>25</td>
<td>3</td>
<td>PCA</td>
<td>NO</td>
</tr>
<tr>
<td>[94]</td>
<td>40</td>
<td>2</td>
<td>consistent flow bundle</td>
<td>NO</td>
</tr>
<tr>
<td>[82]</td>
<td>4</td>
<td>2</td>
<td>frequency domain</td>
<td>NO</td>
</tr>
<tr>
<td>[34]</td>
<td>100</td>
<td>2</td>
<td>assume known motion</td>
<td>Kalman filter</td>
</tr>
<tr>
<td>[36, 35]</td>
<td>30</td>
<td>3</td>
<td>hierarchical estimates</td>
<td>bilateral-TV</td>
</tr>
<tr>
<td>[75]*</td>
<td>15, 60</td>
<td>2</td>
<td>SIFT+RANSAC</td>
<td>NO</td>
</tr>
<tr>
<td>[89]</td>
<td>20</td>
<td>4</td>
<td>SIFT+RANSAC</td>
<td>Least-square</td>
</tr>
<tr>
<td>[5]</td>
<td>10</td>
<td>2</td>
<td>region tracking</td>
<td>motion analysis</td>
</tr>
<tr>
<td>[3]</td>
<td>20, 40</td>
<td>8</td>
<td>moment-based or</td>
<td>Wiener or</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Harris + RANSAC</td>
<td>MRNSD [67]</td>
</tr>
<tr>
<td>[32]</td>
<td>1</td>
<td>2</td>
<td>implicit: NLM</td>
<td>NO</td>
</tr>
<tr>
<td>[31]</td>
<td>20</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[74]</td>
<td>30</td>
<td>3</td>
<td>implicit: NLM</td>
<td>TV</td>
</tr>
<tr>
<td>[78]</td>
<td></td>
<td></td>
<td>kernel regression</td>
<td>bilateral-TV</td>
</tr>
<tr>
<td>[24]</td>
<td>9</td>
<td>3</td>
<td>Video-BM3D</td>
<td>zero-padding spectra</td>
</tr>
</tbody>
</table>

Table 2: Multi-image SIFT for registration

<table>
<thead>
<tr>
<th>Application</th>
<th># of images</th>
<th>Registration</th>
<th>Blending method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]*</td>
<td>manuscript</td>
<td>Not mention</td>
<td>SIFT + RANSAC</td>
</tr>
<tr>
<td>[64]</td>
<td>registration</td>
<td>30 ultrasound</td>
<td>SIFT + threshold +</td>
</tr>
<tr>
<td></td>
<td>60 MRI</td>
<td>least-square for affine</td>
<td></td>
</tr>
<tr>
<td>[84]</td>
<td>Mosaic</td>
<td>200</td>
<td>SIFT + RANSAC</td>
</tr>
<tr>
<td>[46]</td>
<td></td>
<td>10</td>
<td>SIFT + RANSAC</td>
</tr>
<tr>
<td>[52]</td>
<td>stitching</td>
<td>6</td>
<td>SIFT + RANSAC</td>
</tr>
<tr>
<td>[93]</td>
<td>head tracking</td>
<td>1020</td>
<td>SIFT + RANSAC</td>
</tr>
</tbody>
</table>
in the derivative image. In order to minimize the effect of edges small windows were preferred, with $3 \times 3$ or $5 \times 5$ pixels. The sample variance of small patches or the pointwise derivatives provide a non robust measure and require a considerable number of samples with few outliers to guarantee the correct selection of the standard deviation. We observed that the opposite point of view, that is, the use of larger windows $15 \times 15$ pixels to $21 \times 21$ pixels permits a more robust estimation. However, since larger windows may contain more edges a much smaller percentile will be preferred to the median, in practice the 1% or the 0.5%.

Noise in real photograph images is signal dependent. In order to adapt the noise estimation strategies, the gray level image histogram will be divided adaptively into a fixed number of bins having all the same number of samples. This is preferable to classical approaches where the gray range is divided into equal intervals. Such a uniform division can cause many bins to be almost empty.

To evaluate if a signal dependent noise can be estimated from a single image, 110 images were taken with a Nikon D80, with ISO 100 and very good illumination conditions. These are the best conditions we can expect to have a low noise standard deviation. These color images were converted to gray level by averaging the three color values at each pixel. Finally factor 3 sub-sampling was applied by averaging square groups of nine pixels. These operations having divided the noise standard deviation by slightly more than five, these images can be considered as noise free. Finally, a signal dependent noise was added to them, with variance $8 + 2u$ where $u$ was the noiseless grey level.

The uniform and adaptive divisions of the grey level range in a fixed number of 15 bins were compared, and several noise estimation methods were applied to estimate the noise standard deviation inside each bin. The performance of all methods are compared in Table 3 showing the average and standard deviation of the errors between the estimated and original noise curves. The best estimate is obtained by applying the proposed strategy using the variance of large patches rather than small ones or point derivatives. These measurements also confirm that the division of the grey level range into bins with fixed cardinality is preferable to the fixed length interval division. This experiment confirms that a signal dependent noise can be estimated with a high accuracy.

**Ground truth?** In order to evaluate the performance of such a noise estimation algorithm in real images we need a ground truth to compare with. This ground truth can be obtained for a given camera by taking a sequence of images of the same pattern, after fixing the camera on a pedestal. All camera parameters remain unchanged for all photographs of the sequence, thus avoiding different exposure times or apertures. The temporal average and standard deviation of the whole sequence of images can therefore be computed without any further registration. The use of a piecewise constant image reduces the effect of small vibrations of the camera, see Fig. 2. The noise in each channel is estimated independently. Each color range is divided adaptively into a fixed number of bins taking into account the color channel histogram. Inside each bin a percentile is used to estimate the standard deviation.

Fig. 3 displays the ground truth estimated curves with this strategy, both in RAW and JPEG format for two different ISO settings. The ground truth curves are compared with the ones estimated in the first image of the sequence by the proposed single image noise estimation algorithm. For the RAW case, the single image and ground truth estimated curves are nearly identical. Fig. 2 shows a lack of red in the RAW image of the calibration pattern, even if this pattern is actually gray. This effect is corrected by the white balance as observed in the JPEG image.
Table 3: A signal dependent noise with variance $8 + 2u$ is added to 110 noise free images. The uniform and adaptive strategies for dividing the grey level range in a fixed number of 15 bins are compared. For each strategy, the following noise estimation methods in each bin are compared: median of absolute derivatives (MAD), robust median of absolute derivatives (RMAD), median of sample variance of patches $3 \times 3$ of the derivative image (MVPD) and 0.005 percentile of sample variance of patches $21 \times 21$ of the derivative image (MVPD2). Are displayed the average and standard deviation of the errors between the estimated and original noise curves for the 110 images.

The ground truth noise curves estimated from the JPEG images do not agree at all with the classical noise model. This is due to the various image range nonlinear transformations applied by the camera hardware during the image formation, which modify the nature and standard deviation of the noise. The ground truth and single image estimated curves in the JPEG case have a similar shape but a different magnitude. The main new feature is that the interpolation and low pass filtering applied to the originally measured values have strongly altered the high frequency components of the noise. Thus, the noise statistics are no longer computable from a local patch of the image. The estimation of such a noise curve can only be accomplished by computing the temporal variance in a sequence of images of the same scene.

4.2 Multi-image noise estimation

A temporal average requires the images of the sequence to be perfectly registered. Yet, this registration rises a serious technical objection: how to register globally the images of a burst? Fortunately, there are several situations where the series of snapshots are
Figure 3: Ground truth and single image noise estimates for the RAW and JPEG images of Fig. 2. The estimated curve by the temporal average and standard deviation coincide with the one estimated from the first image by the proposed single image noise estimation algorithm. This is not the case for the JPEG images. The ground truth and single image estimated curves in the JPEG case have a similar shape but a different magnitude. The interpolation and low pass filtering applied to the original measured values have altered the high frequency components of the noise and have correlated its low frequencies. This means that the noise statistics are no longer computable from a local patch of the image. The estimation of a noise curve can only be accomplished by computing the temporal variance in a sequence of images of the same scene.

indeed related to each other by a homography, and we shall explore these situations first. The homography assumption is actually valid in any of the following situations:

1. the only motion of the camera is an arbitrary rotation around its optic center;
2. the photographed objects share the same plane in the 3D scene;
3. the whole scene is far away from the camera.

The computation of an homography between a pair of images needs the accurate correspondence of at least four points in each image. Finding key points in images and matching them is a fundamental step for many computer vision and image processing applications. One of the most robust is the Scale Invariant Feature Transform (SIFT) [55], which we will use. Other possible methods allowing for large baselines are [61, 62, 60, 71, 66, 65], but we are here using images taken with only slight changes of view point.

Because wrong matches occur in the SIFT method, an accurate estimate of the dominant homography will require the elimination of outliers. The standard method to eliminate outliers is RANSAC (RANdom SAmple Consensus) [38]. However, it is efficient only when outliers are a small portion of the whole matching set. For this reason several variants have been proposed to improve the performance of outlier elimination, the principal being [79, 90, 81, 68, 63]. The main difference between our approach and the classic outlier elimination is the fact that we dispose of a whole sequence of images and not just of a pair. Instead of choosing a more elaborate version than RANSAC, we preferred to exploit the sequence redundancy in order to improve the registration stage.
The goal is to estimate a dominant homography for the whole set of images, which are typically a few dozens. Only matches which are common to the whole sequence must be kept. In other terms, the keypoints of the first image are kept only if they are matched with another keypoint in any other image of the sequence. This constraint eliminates most of the outliers (see Algorithm 1). In order to apply such a strategy, we assume that the images overlap considerably. Recall that the purpose is not to make a mosaic or a panorama, but to estimate the noise curve and eventually to denoise the sequence.

A temporal average and standard deviation is computed for the registered sequence. The average values are used to build a histogram and to divide the grey level range adaptively. Inside each bin, the median value of the corresponding standard deviations is taken.

Algorithm 1 Hybrid Accumulation After Registration Algorithm

**Input** Initial set of images $I_0, I_1, \cdots, I_n$, obtained from a burst

**SIFT**
Apply the SIFT algorithm between to each pair $(I_0, I_j)$, $j = 1, \cdots, n$. Call $S_j$ the set of matches.
Retain from $S_j$ only the matches for which the matching keypoint in $I_0$ has a match in all other images.

**RANSAC**
Set number of agreed points, $m$, to 0.
while the number of trials does not exceed $N$

Pick up 4 random points from $S_0$
for (each $j > 0$) do

Compute the homography using these 4 points and the corresponding ones in $S_j$
Add to $m$ the number of points in $S_j$ which agree with this homography up to the precision $p$.
end for
If $m > \text{maxim}$, then $\text{maxim} = m$ and save the set of agreed points in the whole sequence
end while
Compute for each pair, the homography $H_j$ with the selected points.

**FUSION**
Apply the homography $H_j$ to each image obtaining $\bar{I}_j$, $j = 1, \cdots, n$.
Average the transformed images obtaining the mean $\mu(x,y)$. Compute also $\sigma(x,y)$, the temporal standard deviation.
Estimate the noise curve using $\sigma(x,y)$, getting $\sigma_n(u)$ the standard deviation associated to each color $u$.
Obtain the final estimate:

$$(1 - w(\mu,\sigma))\mu(x,y) + w(\mu,\sigma) \text{NL}(I_0)(x,y),$$

where NL is the NL-means algorithm (Buades et al. [15]) and the function $w(\nu,\sigma)$ is defined by

$$w(\nu,\sigma) = \begin{cases} 
0 & \text{if } \sigma < 1.5\sigma_n(\mu) \\
\frac{\sigma-1.5\sigma_n(\mu)}{1.5\sigma_n(\mu)} & \text{if } 1.5\sigma_n(\mu) < \sigma < 3\sigma_n(\mu) \\
1 & \text{if } \sigma > 3\sigma_n(\mu) 
\end{cases}$$
Fig. 4 displays three frames from an image sequence with a rotating pattern and a fixed pedestal. The noise curves estimated from the first image with the single image algorithm and those from the registered and averaged sequence are displayed in the same figure. The estimated curves in the raw image coincide if either of both strategies is applied. However, as previously observed these are quite different when we take into account the JPEG image.

Images taken with indoor lights often show fast variations of the contrast and brightness, like those in Fig. 5. This brightness must be rendered consistent through all the images, so that the standard deviation along time is due to the noise essentially and not to the changes of lights. For this reason, a joint histogram equalization must conservatively be applied before the noise estimation chain. The Midway equalization method proposed in [27, 28] is the ideal tool to do so, since it forces all images to adopt a joint
Figure 5: Top: two frames of an image sequence with variations of brightness. Noise curve estimated by temporal average and standard deviation after registration. Bottom: the same two frames of the sequence after a joint histogram equalization [28] and estimated noise curves. The second estimation is correct. The first was not, because of the almost imperceptible lighting conditions.

midway histogram which is indeed a kind of barycenter of the histograms of all images in the burst. Fig. 5 illustrates the noise estimation after and before color equalization.

5 Average after registration denoising

The core idea of the average after registration (AAR) denoising method is that the various values at a cross-registered pixels obtained by a burst are i.i.d.. Thus, averaging the registered images amounts to averaging several realizations of these random variables. An easy calculation shows that this increases the SNR by a factor proportional to $\sqrt{n}$, where $n$ is the number of shots in the burst.

There is a strong argument in favor of denoising by simple averaging of the registered samples instead of block-matching strategies. If a fine non-periodic texture is present in an image, it is virtually indistinguishable from noise, and actually contains a flat spectrum part which has the same Fourier spectrum as the white noise. Such fine textures can be distinguished from noise only if several samples of the same texture are present in other frames and can be accurately registered. Now, state of the art denoising methods (e.g. BM3D) are based on nonlocal block matching, which is at risk to confound the repeated noise-like textures with real noise. A registration process which is far more global than block matching, using strong features elsewhere in the image, should permit a safer denoising by accumulation, provided the registration is sub-pixel accurate and the number of images sufficient.

A simple test illustrates this superior noise reduction and texture preservation on fine non periodic textures. A image was randomly translated by non integer shifts, and signal dependent noise was added to yield an image sequence of sixteen noisy images. Figure 6 shows the first image of the sequence and its denoised version obtained by accumulation after registration (AAR). The theoretical noise reduction factor with sixteen images is
four. This factor is indeed reached by the accumulation process. Table 4 displays the mean square error between the original image and the denoised one by the different methods. Block based algorithms such as NLmeans [15] and BM3D [23], have a considerably larger error, even if their noise reduction could be theoretically superior due to their two dimensional averaging support. But fine details are lost in the local comparison of small image blocks.

![Figure 6: Noise curve. From top to bottom: one of the simulated images by moving the image and adding Poisson noise, denoised by accumulation after registration and the noise curve obtained by the accumulation process using the sixteen images. The standard deviation of the noise (Y-axis) fits to the square root of the intensity (X-axis).](image)

As mentioned in the introduction, the registration by using the SIFT algorithm and computing a homography registration is by now a standard approach in the image fusion literature. The main difference of the proposed approach with anterior work is that the mentioned works do not account for registration errors. Yet, in general, the images of a 3D scene are not related by a homography, but by an epipolar geometry [43]. Even if the camera is well-calibrated, a 3D point-to-point correspondence is impossible to obtain without computing the depth of the 3D scene. However, as we mentioned, a camera held steadily in the hand theoretically produces images deduced from each other by a homography, the principal image motion being due to slight rotations of the camera. Nonetheless, we should not expect that a simple homography will be perfectly accurate everywhere in each pair, but only on a significant part. A coherent registration will be obtained by retaining only the SIFT matches that are common to the whole burst. Therefore the registration applies a joint RANSAC strategy, as exposed in Algorithm 1. This ensures that the same background objects are used in all images to compute the corresponding homographies.
The main new feature of the algorithm is this: The averaging is applied only at pixels where the observed standard deviation after registration is close to the one predicted by the estimated noise model. Thus, there is no risk whatsoever associated with AAR, because it only averages sets of samples whose variability is noise compatible.

At the other pixels, the conservative strategy is to apply a state of the art video denoising algorithm such as the spatiotemporal NL-means algorithm or BM3D. To obtain a smooth transition between the averaged pixels and the NL-means denoised pixels, a weighting function is used. This function is equal to 0 when the standard deviation of the current pixel is lower than 1.5 times the estimated noise standard deviation, and equal to 1 if it is larger than 3 times the estimated noise standard deviation. The weights are linearly interpolated between 1.5 and 3.

6 Discussion and experimentation

We will compare the visual quality of restored images from real burst sequences. The focus is on JPEG images, which usually contain non white noise and color artifacts. As we illustrated in the previous sections, the variability of the color at a certain pixel cannot be estimated from a single image but from a whole sequence. We will compare the denoised images by using AAR as well as the classical block based denoising algorithms, NL-means. Fig. 7 shows the results obtained on three different bursts. Each experiment shows in turn: a) three images extracted from the burst, b) the burst average after registration performed at all points, followed by a mask of the image regions in which the temporal standard deviation is significantly larger than the standard deviation predicted by the noise estimate. At all of these points a block based denoising estimate is used instead of the temporal mean. The final combined image, obtained by an hybridization of the average registration and NL-Means or BM3D, is the right image in each second row.

The first experimental data was provided by the company DxO Labs. It captures a rotating pattern with a fixed pedestal. In this case, the dominant homography is a rotation of the main circular pattern, which contains more SIFT points than the pedestal region. Since the proposed algorithm only finds a dominant homography, which is the rotation of the pattern, the simple average fails to denoise the region of the fixed pedestals and of the uniform background. As shown in the white parts of the mask, these regions are detected because they have an excessive temporal standard deviation. They are therefore treated by NL-means or BM3D in the final hybrid result. The whole pattern itself is restored by pure average after registration.

The second burst consists of two books, a newspaper and a moving mouse. Since the
dominant homography is computed on still parts, the books and the background, the moving mouse is totally blurred by the averaging after registration, while the rest of the scene is correctly fused. As a consequence, AAR uses the average everywhere, except the part swept by the mouse.

The last burst is a sequence of photographs with short exposure time of a large painting taken in Musée d’Orsay, Martyrs chrétiens entrant lamphithéâtre by Léon Bénouville. Making good photographs of paintings in the dim light of most museums is a good direct application for the proposed algorithm, since the images of the painting are related by a homography even with large changes of view point, the painting being flat. As a result, the average is everywhere favored by the hybrid scheme. Details on the restored images and comparison with BM3D are shown in Fig. 8-10. Dim light images are displayed after their color values have been stretched to $[0, 255]$.

References


Figure 7: In each double row: three images of a sequence in the first row. In the second row on the left the average after registration, in the middle the mask of points with a too large temporal standard deviation, and on the right the restored image by hybrid method. These experiments illustrate how the hybrid method detects and corrects the potential wrong registrations due to local errors in the global homography.
Figure 8: Detail from image in Fig. 7. From left to right: original image, NL-means (BM3D gives a similar result) and hybrid AAR. The images may need to be zoomed in on a screen to compare details and textures. Compare the fine texture details in the trees and the noise in the sky.

Figure 9: Detail from image in Fig. 7. From left to right: original image, BM3D (considered the best state of the art video denoiser) and AAR. The images are displayed after their color values have been stretched to [0, 255]. The images may need to be zoomed in on a screen to compare details and textures. Notice how large color spots due to the demosaicking and to JPEG have been corrected in the final result.

Figure 10: Detail from image in Fig. 7. From left to right: original image, BM3D (considered the best state of the art video denoiser) and AAR. Images are displayed after their color values have been stretched to [0, 255]. The images may need to be zoomed in on a screen to compare details and textures. Compare details on the face and on the wall texture.
Figure 11: Top: initial image of the burst containing six images. Bottom: details on the initial and hybrid AAR images.