Domain Decomposition Algorithms for Real-Time Homogeneous Diffusion Inpainting in 4K

Niklas Kämper and Joachim Weickert

Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Campus E1.7, Saarland University, 66041 Saarbrücken, Germany.
{kaemper, weickert}@mia.uni-saarland.de

Abstract

Inpainting-based compression methods are qualitatively promising alternatives to transform-based codecs, but they suffer from the high computational cost of the inpainting step. This prevents them from being applicable to time-critical scenarios such as real-time inpainting of 4K images. As a remedy, we adapt state-of-the-art numerical algorithms of domain decomposition type to this problem. They decompose the image domain into multiple overlapping blocks that can be inpainted in parallel by means of modern GPUs. In contrast to classical block decompositions such as the ones in JPEG, the global inpainting problem is solved without creating block artefacts. We consider the popular homogeneous diffusion inpainting and supplement it with a multilevel version of an optimised restricted additive Schwarz (ORAS) method that solves the local problems with a conjugate gradient algorithm. This enables us to perform real-time inpainting of 4K colour images on contemporary GPUs, which is substantially more efficient than previous algorithms for diffusion-based inpainting.

Index Terms— Inpainting, Homogeneous Diffusion, Domain Decomposition, Restricted Additive Schwarz Method.

1. Introduction

Inpainting-based image compression methods can be competitive alternatives to classical lossy transform-based image codecs. They only store the values of a few carefully selected pixels. In the decoding phase, they reconstruct the missing image parts by inpainting. Gall et al. [1] introduced nonlinear diffusion-based inpainting for image compression in 2005. Schmaltz et al. [2] improved this idea and Peter et al. [3] extended it to colour images. They showed that they can outperform JPEG [4] and JPEG2000 [5] qualitatively for real-world test images with small to medium amount of texture.

Interestingly, already a simple linear process such as homogeneous diffusion inpainting [6] can be a powerful component, provided that the locations and function values of the sparse inpainting data are carefully optimised [4]. Figure 2 shows an example. For certain image types such as cartoon-like images [8], depth maps [9], and flow fields [10] it can even produce state-of-the-art results. Homogeneous diffusion inpainting offers the advantage of being parameter-free, and its discretisation leads to linear systems of equations (in contrast to nonlinear diffusion which creates nonlinear systems). Each unknown represents a greyscale or colour channel value of a pixel that is to be inpainted.

Unfortunately, such linear systems are fairly large, and computing their numerical solution can be time-consuming. This is the reason why inpainting-based compression is usually slower than transform-based codecs and considered to be too slow for time-critical applications. Nowadays, 4K colour images of size $3840 \times 2160$ pixels constitute a standard for TV applications, and it would be desirable to achieve real-time decoding with at least 30 frames per second. In spite of substantial research to accelerate inpainting-based compression [11, 8, 12, 13, 14, 15] this is not possible so far: Current approaches lack behind these requirements by at least one order of magnitude. However, most of their numerical solvers have been tailored towards sequential or mildly parallel (in the sense of multi-core CPUs) architectures. They do not exploit the potential of dedicated algorithms for highly parallel GPUs that are widely available these days.

1.1. Our Contribution

The goal of our present paper is to address this problem by advocating and adapting a class of powerful numerical algorithms: domain decomposition methods [17, 18]. Apart from a few exceptions such as [19, 20, 21], they are hardly used in image processing so far. Domain decomposition algorithms subdivide the image domain into multiple subdomains and solve the linear systems on each subdomain in parallel. Firstly, this reduces the overall computational load for solvers with complexity worse than linear. Secondly, this decoupling is very well-suited for highly parallel architectures such as GPUs. By permitting some communication across subdomain boundaries and iterating this concept, one encourages convergence to the exact solution of the global problem. Thus, no artefacts at subdomain boundaries arise. This is a decisive advantage over widely-used block decompositions in image processing, such as the $8 \times 8$ pixel partitions in JPEG. They suffer from visible artefacts. We show that by developing an adapted multilevel domain decomposition method for homogeneous diffusion inpainting on a contemporary GPU, we can inpaint 4K colour images in real-time.

1.2. Related Work

Let us now review earlier approaches to accelerate diffusion inpainting. Multigrid methods [22, 23] belong to the most efficient numerical solvers for linear and nonlinear systems on sequential architectures. Both Köstler et al. [11] and Mainberger et al. [8] use them for homogeneous diffusion inpainting. They also consider mildly parallel architectures such as multicore CPUs [6] and the Playstation 3 hardware [11].

An approach by Hoffmann et al. [13] is based on Green’s functions. It has the advantage that its runtime depends on the number of
stored pixels instead of the overall number of pixels. It can outperform multigrid for very sparse inpainting data.

Chizhov and Weickert [14] consider adaptive finite element approximations instead of finite difference discretisations. This can lead to faster inpaintings, since one replaces a fine regular pixel grid by a coarser adaptive triangulation with less unknowns.

For the more sophisticated anisotropic nonlinear diffusion, Peter et al. [12] achieved real-time decoding of 640 × 480 videos on an Nvidia GeForce GTX 460 GPU. It relies on accelerated explicit finite difference schemes [24] that are well-suited for parallelisation and benefit strongly from a good initialisation from the previous frame. This advantage would be unavailable for individual inpaintings of unrelated images.

Two other real-time video players that exploit temporal coherence go back to Andris et al. [15, 16]. They combine global homogeneous diffusion inpainting of keyframes with optic flow based prediction of interframes. The recent paper [16] reports real-time performance for FullHD colour videos on a multicore CPU. Its inpainting involves a multilevel conjugate gradient method [25]. This discussion shows what distinguishes our work from previous papers: Its highly parallel nature fully exploits the performance for FullHD colour videos on a multicore CPU. Its inpainting an RGB colour image leads to three linear systems of this type that yield inpaintings with reflecting boundary conditions.

For digital images, we discretise [4] with finite differences and obtain a linear system of equations [26, 8]. Its solution specifies the reconstructed grey values in all pixels of the inpainting domain. More specifically, let \( f \in \mathbb{R}^N \) be a discretised version of \( f \) with \( N \) pixels. The pixels locations inside \( K \) constitute the so-called inpainting mask. The confidence function \( c \) is replaced by a diagonal matrix \( C \in \mathbb{R}^{N \times N} \). Its diagonal entries are 1 in mask pixels and 0 elsewhere. Then the discrete counterpart of (4) is given by

\[
C \cdot (u - f) - (I - C) \cdot L u = 0
\]

where \( I \in \mathbb{R}^{N \times N} \) denotes the identity matrix, and \( L \in \mathbb{R}^{N \times N} \) represents the discrete Laplacian with reflecting boundary conditions. We can rewrite (5) as a linear system of equations

\[
Au = b
\]

with \( A = C - (I - C) L \) and \( b = Cf \). Inpainting an RGB colour image leads to three linear systems of this type that yield inpaintings of all three channels. To solve such systems efficiently on parallel hardware, let us now discuss a specific domain decomposition technique: the restricted additive Schwarz method.

### 3. Restricted Additive Schwarz Method

The restricted additive Schwarz (RAS) method [27] is an iterative technique for solving the linear system (6). It is one of the simplest domain decomposition methods and easy to parallelise. First the image domain \( \Omega \) with \( N \) pixels is partitioned into \( k \) overlapping subdomains \( \Omega_1, \ldots, \Omega_k \subset \Omega \) such that \( \bigcup_{k}^i \Omega_i = \Omega \). Figure 1 shows an example of a subdivision into four overlapping blocks. Let \( |\Omega_i| \) denote the number of pixels in \( \Omega_i \). In iteration step \( n \), we compute local corrections \( u_i \in [0, \ldots, 1] \) on every subdomain \( \Omega_i \) by solving multiple smaller systems of equations. They are given by

\[
R_i A R_i^T u_i = R_i r^n,
\]

where \( r^n = b - Au^n \) is the residual from the previous iteration. The upper index \( n \) denotes the iteration number and not a power. \( R_i \in \mathbb{R}^{N \times |\Omega_i|} \) is a restriction matrix that restricts vectors \( u \in \mathbb{R}^N \) on the global domain \( \Omega \) to local vectors \( u_i \in \mathbb{R}^{|\Omega_i|} \) on the subdomain \( \Omega_i \). It is defined as

\[
(R_i)_{\ell,k} = \begin{cases} 
1 & \text{if } \ell = k \text{ and } \ell \in \Omega_i, \\
0 & \text{else}.
\end{cases}
\]
5. EXPERIMENTS

Let us now evaluate our domain decomposition method w.r.t. runtime. We consider two variants: a single level and a multilevel version with three resolution levels. For both variants we use a parallelised GPU implementation. We compare our ORAS approach to a conjugate gradient (CG) method because it is easy to parallelise and to implement efficiently on a GPU. Using a multilevel CG method allows a fair comparison to our multilevel ORAS technique. Moreover, the multilevel CG approach is the core algorithm in the recent inpainting-based video player of Andris et al. [16] which achieved real-time performance in FullHD resolution on a multicore CPU. Thus, we can judge where we stand w.r.t. the current state-of-the-art. It must be emphasised that other fast solvers such as multigrid methods or the Green’s function approach are less suited for GPU implementations. Thus, they are excluded in our comparison. All experiments were conducted on an AMD Ryzen 5900X@3.7GHz with a Nvidia GeForce GTX 1080 Ti GPU. We tested our method on the 4K image lofsdalen shown in Figure 2(a), which has a resolution of $3840 \times 2160$. This image is a typical real-world image with a variety of coarse and fine structures and a medium amount of texture. Therefore, it is a good representative for analysing the performance of our method. The optimised inpainting mask was obtained by a Voronoi densification [14] and has a data density of 5%. The corresponding inpainting solution can be seen in Figure 2(c).

5.1. Timing Results

Figure 5(a) shows the residual decay of the evaluated algorithms as a function of the runtime. As expected we observe that the multilevel variants of the CG and ORAS methods are substantially more efficient than their single level counterparts. More importantly, we also see that our multilevel ORAS approach is more than four times faster than multilevel CG. This demonstrates the superiority of domain decomposition strategies over simpler algorithms that are equally well parallelisable.
Figure 3(b) displays the improvement of the peak signal-to-noise ratio (PSNR) with decreasing relative residual norm. For colour images, we base the PSNR on the mean square error averaged over the three channels. One can observe that each method achieves the optimal PSNR at a different relative residual norm. This demonstrates that the relative residual norm by itself is not a suitable stopping criterion and that one should look at the PSNR in order to determine which relative residual norm is sufficient for each method. The figure shows that our multilevel ORAS algorithm reaches the optimal PSNR already at a substantially larger relative residual compared to the competing methods. A relative residual norm of $10^{-3}$ is sufficient for our multilevel ORAS algorithm and obtained after four global iterations which requires 25.2 milliseconds. In contrast, the competing multilevel CG solver needs 25 iterations and 115.5 milliseconds to reach the same relative residual, resulting in a speed-up factor of 4.58. However, to achieve the optimal PSNR, it needs at least a relative residual smaller than $5 \cdot 10^{-4}$, which is obtained after 141.5 milliseconds. Thus, domain decomposition offers a speed-up factor of 5.61.

5.2. Scaling Results

In order to evaluate the performance of our inpainting method over different image sizes, we conducted an experiment over resolutions ranging from $960 \times 540$ to $3840 \times 2160$. The results are shown in Figure 4. We see that for all image resolutions, our method is at least four times faster than the multilevel CG technique. The CG approach can inpaint a FullHD image in 33.3 milliseconds, which corresponds to 30 frames per second. However, it is not able to inpaint higher resolution images in real-time. On the other hand, our method can perform real-time inpainting on 4K images with more than 30 frames per second.

Figure 4 also reveals that both methods show a nearly linear behaviour in the double logarithmic plot. This demonstrates the presence of an underlying power law. Its power is given by the slope of the line, which is approximately 1. Thus, we observe an ideal scaling behaviour where the computational time grows linearly with the number of pixels. This suggests that the multilevel ORAS approach offers the best of two worlds: a linear scaling behaviour that is characteristic for full multigrid methods on model problems [23], and a perfect suitability for parallel architectures which originates from the domain decomposition.

6. CONCLUSIONS AND OUTLOOK

We have seen that it pays off to marry state-of-the-art numerical algorithms with promising ideas from image processing and the computing power of modern parallel hardware. This enabled us for the first time to perform diffusion-based sparse inpainting of 4K images in real-time on contemporary GPUs. In view of this success, it is surprising that domain decomposition ideas have hardly been explored in image analysis so far. Our results also demonstrate that inpainting-based compression has left its infancy to become a serious alternative to classical transform-based codecs not only in terms of quality, but also for time-critical applications.

In our ongoing work, we are extending our domain decomposition framework to more advanced inpainting operators that may offer further quality improvements.
7. REFERENCES

[1] I. Galić, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.-P. Seidel, “Towards PDE-based image compression,” in Variational, Geometric and Level-Set Methods in Computer Vision, N. Paragios, O. Faugeras, T. Chan, and C. Schnörr, Eds., vol. 3752 of Lecture Notes in Computer Science, pp. 37–48. Springer, Berlin, 2005.

[2] C. Schmaltz, P. Peter, M. Mainberger, F. Ebel, J. Weickert, and A. Bruhn, “Understanding, optimising, and extending data compression with anisotropic diffusion,” International Journal of Computer Vision, vol. 108, no. 3, pp. 222–240, July 2014.

[3] P. Peter, L. Kaufhold, and J. Weickert, “Turning diffusion-based image colorization into efficient color compression,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 860–869, Feb. 2017.

[4] W. B. Pennebaker and J. L. Mitchell, JPEG: Still Image Data Compression Standard, Springer, New York, 1992.

[5] D. S. Taubman and M. W. Marcellin, JPEG 2000: Image Compression Fundamentals, Standards and Practice, Kluwer, Boston, 2002.

[6] S. Carlsson, “Sketch based coding of grey level images,” Signal Processing, vol. 15, pp. 57–83, 1988.

[7] M. Mainberger, S. Hoffmann, J. Weickert, C. H. Tang, D. Johannsen, F. Neumann, and B. Doerr, “Optimising spatial and tonal data for homogeneous diffusion inpainting,” in Scale Space and Variational Methods in Computer Vision, A. M. Bruckstein, B. ter Haar Romeny, A. M. Bronstein, and M. M. Bronstein, Eds., vol. 6667 of Lecture Notes in Computer Science, pp. 26–37. Springer, Berlin, 2012.

[8] M. Mainberger, A. Bruhn, J. Weickert, and S. Forchhammer, “Edge-based compression of cartoon-like images with homogeneous diffusion,” Pattern Recognition, vol. 44, no. 9, pp. 1859–1873, Sept. 2011.

[9] S. Hoffmann, M. Mainberger, J. Weickert, and M. Puhl, “Compression of depth maps with segment-based homogeneous diffusion,” in Scale Space and Variational Methods in Computer Vision, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds., vol. 7893 of Lecture Notes in Computer Science, pp. 319–330. Springer, Berlin, 2013.

[10] F. Jost, P. Peter, and J. Weickert, “Compressing flow fields with edge-aware homogeneous diffusion inpainting,” in Proc. 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, May 2020, pp. 2198–2202, IEEE Computer Society Press.

[11] H. Köstler, M. Sturmer, C. Freundl, and U. Rüde, “PDE based video compression in real time,” Tech. Rep. 07-11, Lehrstuhl für Informatik 10, Univ. Erlangen–Nürnberg, Germany, 2007.

[12] P. Peter, C. Schmaltz, N. Mach, M. Mainberger, and J. Weickert, “Beyond pure quality: Progressive mode, region of interest coding and real time video decoding in PDE-based image compression,” Journal of Visual Communication and Image Representation, vol. 31, pp. 256–265, Aug. 2015.

[13] S. Hoffmann, G. Plonka, and J. Weickert, “Discrete Green’s functions for harmonic and biharmonic inpainting with sparse atoms,” in Energy Minimization Methods in Computer Vision and Pattern Recognition, X.-C. Tai, E. Bae, T. F. Chan, and M. Lysaker, Eds., vol. 8932 of Lecture Notes in Computer Science, pp. 169–182. Springer, Berlin, 2015.

[14] V. Chizhov and J. Weickert, “Efficient data optimisation for harmonic inpainting with finite elements,” in Computer Analysis of Images and Patterns, N. Tsapatsoulis, A. Panayides, T. Theo, A. Lantitis, C. Pattichis, and M. Vento, Eds., vol. 13053 of Lecture Notes in Computer Science, pp. 432–441. Springer, Cham, 2021.

[15] S. Andris, P. Peter, and J. Weickert, “A proof-of-concept framework for PDE-based video compression,” in Proc. 32nd Picture Coding Symposium, Nuremberg, Germany, Dec. 2016.

[16] S. Andris, P. Peter, R. M. Kaja Mohideen, J. Weickert, and S. Hoffmann, “Inpainting-based video compression in FullHD,” in Scale Space and Variational Methods in Computer Vision, A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, Eds., vol. 12679 of Lecture Notes in Computer Science, pp. 425–436. Springer, Cham, 2021.

[17] B. Smith, P. Bjørstad, and W. Gropp, Domain Decomposition: Parallel Multilevel Methods for the Solution of Elliptic Partial Differential Equations, Cambridge University Press, Cambridge, UK, 1996.

[18] V. Dolean, P. Jolivet, and F. Nataf, An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation, SIAM, Philadelphia, 2015.

[19] T. Kohlberger, C. Schnörr, A. Bruhn, and J. Weickert, “Domain decomposition for variational optical flow computation,” IEEE Transactions on Image Processing, vol. 14, no. 8, pp. 1125–1137, Aug. 2005.

[20] J. Xu, X.-C. Tai, and L.-L. Wang, “A two-level domain decomposition method for image restoration,” Inverse Problems and Imaging, vol. 4, no. 3, pp. 523–545, Aug. 2010.

[21] R. Chen, J. Huang, and X.-C. Cai, “A parallel domain decomposition algorithm for large scale image denoising,” Inverse Problems and Imaging, vol. 13, no. 6, pp. 1259–1282, Dec. 2019.

[22] A. Brandt, “Multi-level adaptive solutions to boundary-value problems,” Mathematics of Computation, vol. 31, no. 138, pp. 333–390, Apr. 1977.

[23] W. L. Briggs, V. E. Henson, and S. F. McCormick, A Multigrid Tutorial, SIAM, Philadelphia, second edition, 2000.

[24] J. Weickert, S. Grewenig, C. Schroers, and A. Bruhn, “Cyclic schemes for PDE-based image analysis,” International Journal of Computer Vision, vol. 118, no. 3, pp. 275–299, July 2016.

[25] F. Bornemann and P. Deuflhard, “The cascadic multigrid method for elliptic problems,” Numerische Mathematik, vol. 75, pp. 135–152, 1996.

[26] K. W. Morton and L. M. Mayers, Numerical Solution of Partial Differential Equations, Cambridge University Press, Cambridge, UK, second edition, 2005.

[27] X.-C. Cai and M. Sarkis, “A restricted additive Schwarz preconditioner for general sparse linear systems,” SIAM Journal on Scientific Computing, vol. 21, no. 2, pp. 792–797, 1999.

[28] A. St-Cyr, M. J. Gander, and S. J. Thomas, “Optimized multiplicative, additive, and restricted additive Schwarz preconditioning,” SIAM Journal on Scientific Computing, vol. 29, no. 6, pp. 2402–2425, 2007.

[29] Y. Saad, Iterative Methods for Sparse Linear Systems, SIAM, Philadelphia, second edition, 2003.