Fast operations for a Gabor-frame-based integral equation with equidistant sampling

Chapter 10

Fast operations for a Gabor-frame-based integral equation with equidistant sampling [1]

10.1 Abstract

We consider the computation time of a 3D Gabor-frame based spatial spectral integral equation solver for scattering from dielectric objects embedded in a multi-layer medium. Based on the Gabor frame, a new set of basis functions is proposed, together with a set of equidistant Dirac-delta test functions. Using this construction, we approximate the operations of Fourier transformation and pointwise multiplication by a method that is significantly faster than the original method. A numerical example is included where the computation time is reduced by a factor of 15, while preserving accuracy.

10.2 Introduction

Spatial spectral solvers for computing the scattering from dielectric objects embedded in a multilayered medium are presented in [135, 147, 152] (Chapters 6-8). These methods rely on the Gabor frame and a discretization in both the spatial and the spectral domain. In the spectral domain, a deformation to a complex manifold is employed on which the Green function is smooth enough to allow for a Gabor-frame representation. The main advantage of the Gabor frame is that the Fourier transformation is represented analytically by a simple transposition of the Gabor-coefficient matrix. The downside of the Gabor frame is that the operation of multiplication is represented through an operator that contains a large number of small-sized FFTs and considerable overhead in reordering coefficients [118].

[1] This chapter was submitted as an article for the journal IEEE Antennas and Wireless Propagation Letters [160].

In general, a fast matrix-vector product[2] requires a suitable discretization method. In particular, this method should allow for both a rapid multiplication of functions and a rapid Fourier transformation to and from the spectral domain. A fast multiplication can be achieved when product of basisfunctions are bi-orthogonal to the test functions, since that allows for a coefficient-wise multiplication. Conversely, when they are not bi-orthogognal, each multiplication between two basis functions requires testing with multiple test functions, which is undesirable. The products of Gabor frame functions and the dual Gabor frame functions as test function are not bi-orthogonal. Secondly, a fairly rapid Fourier transformation can be achieved when the test functions are all identical and spaced uniformly, since that allows the use of FFTs. Additionally, the basis functions should also decay sufficiently fast, to allow for a truncation to a small region in the spatial and spectral domain.

We show how the electric field and contrast current density can be represented by a set of basis functions that are related to a Gabor frame and a set of Dirac-delta test functions related to the same Gabor frame that together satisfy the above conditions. In a numerical example, we demonstrate the decrease in computation time and compare results for both discretization methods with comparable accuracy.

10.3 Gabor frame - definitions

We define the Gabor frame [99], starting with the Fourier transform

$$\hat{\varphi}(k) = \int_{-\infty}^{\infty} dx\, \varphi(x) e^{-jkx}. \tag{10.1}$$

For the Gabor frame we follow the definition in [100], i.e.

$$g_{mn}(x) = g(x - m\alpha X) e^{j\beta K n x}, \tag{10.2}$$

with $m, n \in \mathbb{Z}$ and where we use the window function

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}. \tag{10.3}$$

Here, $X = \frac{2\pi}{K}$ is the spacing of the window functions in the spatial domain for an exact frame. In this article, we assume a rational oversampling with $\alpha = \beta = \sqrt{p/q}$, and choose $p = 2$ and $q = 3$. The dual window, $\eta(x)$, is calculated with the aid of the Moore Penrose pseudo-inverse and the method described in [99, 100]. When we have chosen a frame and a dual window, we can calculate the Gabor coefficients of a (square-integrable) function f as

$$f_{mn} = \int_{\mathbb{R}} dx\, \eta_{mn}(x) f(x) \tag{10.4}$$

[2] In the sence of Section 3.4

and function values from Gabor coefficients via

$$f(x) = \sum_{m,n\in\mathbb{Z}} f_{mn} g_{mn}(x). \tag{10.5}$$

In practice, the number of Gabor coefficients is truncated to $m \in \{-M, \cdots, M\}$ and $n \in \{-N, \cdots, N\}$, which yields a total number of $2L + 1$ coefficients. By taking the Fourier transformation of the frame function $g_{mn}(x)$, a spectral frame is defined as

$$\hat{g}_{nm}(k) = \hat{g}(k - n\beta K) e^{j\alpha X m k} e^{-2\pi j \alpha \beta m n}. \tag{10.6}$$

10.4 Basis functions

10.4.1 Representation using lists

In [100], Bastiaans describes the fast Gabor transformation $\mathcal{B}$, that calculates the Gabor coefficients of a function from an uniformly sampled function. This algorithm can also be inverted to obtain $\mathcal{B}^{-1}$, to calculate a list of uniformly sampled function values from a set of Gabor coefficients. The uniformly sampled lists will be denoted in boldface. Since the lists are defined in connection with a particular Gabor frame, the sampling is restricted, i.e. the sampling operator $\mathcal{S}$ samples a function according to that Gabor frame

$$\mathbf{f} = \mathcal{S} \circ f = \{f(\ell \Delta_x), \ell \in -L, \cdots, L\}, \tag{10.7}$$

with L as defined above and where Δ_x depends on the parameters defining the Gabor frame via

$$\Delta_x = \frac{X}{\beta(2N + 1)}. \tag{10.8}$$

10.4.2 Shape of the basis functions

A continuous function is approximated by a set of weighted basis functions. In the context of Gabor frames, the most obvious choice for a basis are the frame functions of Eq. (10.2), which were used in [135, 147, 152, 118]. However, here we will not use the Gabor frame directly as a basis. Instead, we derive the basis functions from the fast Gabor transformation $\mathcal{B}$ of entries to uniformly sampled lists. Since the list representation is tied to the Gabor frame via $\mathcal{B}$, we can compute the Gabor coefficients for each list. The coefficients (in $\ell^2(\mathbb{R})$) are related to a continuous function (in $\mathcal{L}^2(\mathbb{R})$) via Eq. (10.5). Therefore, we define basis functions corresponding to the list $\mathbf{b}_i = \{0, \cdots, 1, \cdots, 0\}$ with a one at position i, that can be found by means of

$$b_i(x) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} g_{mn}(x) \left\{\mathcal{B} \circ \mathbf{b}_i\right\}_{mn}, \tag{10.9}$$

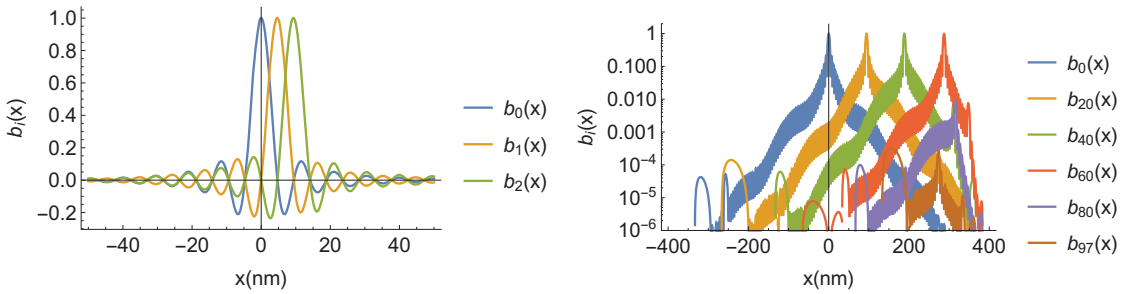

Figure 10.1: Several basis functions for a Gabor frame with $X = 50\text{nm}$, $M = 7$, $N = 6$, $L = 97$.

where the circle denotes the application of an operator.

In Fig. 10.1, we have plotted several of these basis functions. Several interesting features can be observed about these functions. The basis functions have zeroes at the Δ_x grid, except at $x = i\Delta_x$, where they are one for $i < pN/q$. The reason is that they are produced by $\mathcal{B}$ and $\mathcal{B}^{-1}$ on the $\mathbf{b}_i$ list defined for the Δ_x grid. Another observation is that, for high index i, the basis functions are very small. This is caused by the redundancy in the Gabor frame. We also observe that, although these basis functions look similar, they are not one function that is merely shifted in position. There are subtle differences between the basis functions. The final observation that we mention is that $b_i(x)$ resembles a sinc function, which decays slowly. However, outside the simulation domain, $b_i(x)$ decays much faster than the sinc function. In Fig. 10.2(a) and (b), we have also plotted several basis functions in the spectral domain, where it is clearly visible how these basis functions $\hat{b}_i(k_x)$ resemble truncated complex exponentials, since they are produced by functions resembling Dirac-Delta distributions. It is interesting to notice that the truncation has a smooth transition, so the b_i functions rapidly decay to zero at the ends of the simulation domain in the spatial domain.

10.4.3 Testing functions and inner products

In the Gabor-frame discretization we used the dual Gabor frame as test functions, which works well, since the dual Gabor frame is dual to the Gabor frame with respect to the $\mathcal{L}^2(\mathbb{R})$ norm. For the set of $b_i(x)$ basis functions we use Dirac delta test functions on the Δ_x lattice, a set which is dual with respect to the $\mathcal{L}^2(\mathbb{R})$ norm as well.

An $\mathcal{L}^2(\mathbb{R})$-based inner product was employed for the Gabor-frame based method. The computation of an $\mathcal{L}^2$-based inner product was not used here, since all basis functions are slightly different, i.e. they are not simply shifted copies of each other. This means that $\langle b_i|b_j\rangle_{\mathcal{L}^2(\mathbb{R})}$ has a different value for each i and j and there is for example no translation symmetry in the sense that $\langle b_i|b_j\rangle_{\mathcal{L}^2(\mathbb{R})} \neq \langle b_{i+m}|b_{j+m}\rangle_{\mathcal{L}^2(\mathbb{R})}$.

With the Dirac-delta testing procedure, the test functions are $t_i(x) = \delta(x - i\Delta_x)$. As we mentioned before, the basis functions are such that that $\langle b_i|t_j\rangle_{\mathcal{L}^2(\mathbb{R})} = \delta_{ij}$, i.e. they are bi-orthogonal for $i, j < p/qL$. Since $b_i(x)$ and $t_j(x)$ are biorthogonal, we choose the ℓ^2

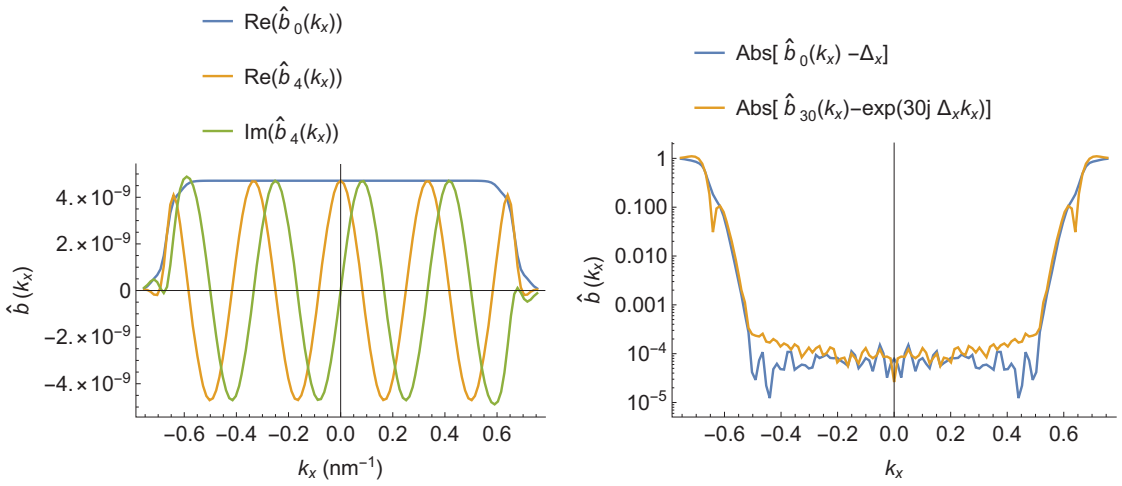

Figure 10.2: (a) The Fourier transform of basis functions of the frame in Figure 10.1. (b) The difference between two spectral basis functions and their corresponding complex exponentials.

inner product

$$\langle \mathbf{f} | \mathbf{h} \rangle = \Delta_x \sum_{i=-L}^{L} f_i \overline{h}_i \approx \langle \mathbf{f} | \mathbf{h} \rangle_{\mathcal{L}^2} = \int_{-\infty}^{\infty} dx f(x) h(x) \tag{10.10}$$

with $\mathbf{f}$ and $\mathbf{h}$ discretized from smooth functions $f(x)$ and $h(x)$, respectively by Eq. (10.7). ¿ Clearly, Eq. (10.10) is equivalent to numerically evaluating the integral in the $\mathcal{L}^2$ inner product between $f(x)$ and $h(x)$ by N equidistant samples. ¿ Therefore, Eq. (10.10) converges to the $\mathcal{L}^2$ inner product when N increases.

10.5 Operations

10.5.1 Multiplication

First, we emphasize that the multiplication operation is non-linear. Consequently, when two functions can be represented well in a Gabor frame, their product is not necessarily well represented in the same Gabor frame. The reason for this is that a spatial multiplication is equivalent to a spectral convolution. For that reason the product of two functions potentially has twice the spectral support of the original functions.

An approximation has to be made to fit the product in the space spanned by the Gabor-frame of the original functions. In the Gabor-frame formulation [118] an (almost) exact multiplication was implemented, but in the end the spectral range is truncated. This is equivalent to an exact multiplication followed by testing with a finite number of test functions.

We apply this procedure to find a multiplication operation with the basis $b_i(x)$, and test functions $t_i(x)$. The procedure for multiplying lists $\mathbf{f}$ and $\mathbf{g}$ then becomes

$$(\mathbf{fg})_i = \int_{x=-\infty}^{\infty} dx \sum_{j,k} \mathbf{f}_j \mathbf{g}_k b_j(x) b_k(x) t_i(x) = \mathbf{f}_i \mathbf{g}_i, \tag{10.11}$$

where we used the property that $b_i(j\Delta_x) = \delta_{ij}$. We would like to emphasize that it is the choice of the test function that yields this simple form of multiplication.

10.5.2 Fourier transformation

The advantage of a uniformly sampled list-based approach is that multiplication is a very fast operation. However, now the Fourier transformation is slower. It is possible to implement the Fourier transformation by succesively applying a fast Gabor transformation $\mathcal{B}$, a Fourier transformation $\mathcal{F}$ on Gabor coefficients and then an inverse Gabor transformation $\hat{\mathcal{B}}^{-1}$, i.e.

$$\hat{\mathbf{f}} = \hat{\mathcal{B}}^{-1} \circ \mathcal{F} \circ \mathcal{B} \circ \mathbf{f}. \tag{10.12}$$

The main drawback is the use of the relatively slow operations $\mathcal{B}$ and $\hat{\mathcal{B}}^{-1}$. For a more optimized method, we exploit the fact that without truncation

$$\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} g_{mn}(x') \eta_{mn}(x) = \delta(x - x'), \tag{10.13}$$

and therefore, using Eq. (4.31), we can write its Fourier transformation in x' as

$$\sum_{m,n=-\infty}^{\infty} \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn} \eta_{mn}(x) = e^{-jk_x x}. \tag{10.14}$$

Now, the Fourier transformation of a function f can be approximated by

$$\begin{aligned} \hat{f}(k_x) &= \int_{-\infty}^{\infty} dx \sum_{m,n=-\infty}^{\infty} f(x) \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn} \eta_{mn}(x) \\ &\approx \Delta_x \sum_{m,n=-\infty}^{\infty} \sum_{\ell=-L}^{\ell=L} f(\ell\Delta_x) \eta_{mn}(\ell\Delta_x) \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn}. \end{aligned} \tag{10.15}$$

In this expression, we recognize the Gabor transformation, Eq. (10.4), as the integral over x, which is replaced by a sum in the second line. This approximation holds when the sampling in Eq. (10.7) approximates f well. This discretized Gabor transformation equals the operator $\mathcal{B}$. The fact that we wrote $\hat{g}_{nm}(k_x)e^{2\pi j\alpha\beta mn}$ instead of $g_{mn}(x)$ exactly represents the Fourier transformation operator $\mathcal{F}$. And finally, the summation over m, n represents the inverse Gabor transformation of Eq. (10.5). The integral in Eq. (10.15) is evaluated on the uniform grid $k_x \in \{-L\Delta_k, \cdots, L\Delta_k\}$ and therefore the m, n summation

equals $\hat{\mathcal{B}}^{-1}$, i.e. the inverse Gabor transformation operator in the spectral domain. Here $\Delta_k = K/\alpha(2N+1)$ is the spectral-domain counterpart of Δ_x. Hence we identified all operators of Eq. (10.12) in Eq. (10.15) and both are equal up to the discretization error in $\mathbf{f}$.

When we apply the discretized version of Eq. (10.14) to the second line of Eq. (10.15), we can write

$$\hat{\mathbf{f}}_m = \sum_{n=-L}^{L} \mathbf{f}_n e^{-j\Delta_x\Delta_k mn} = \sum_{n=-L}^{L} \mathbf{f}_n \exp(-2\pi jmn\frac{p}{q(2L+1)}), \tag{10.16}$$

which looks similar to a discrete Fourier transformation. However, it is the oversampling factor p/q makes the difference.

In case $(2L+1)/p \in \mathbb{N}$, this can be calculated as the Fast Fourier Transform (FFT) of size $q/p(2L+1)$ of $\mathbf{f}^\dagger$ in

$$\mathbf{f}^\dagger_n = \sum_{n=1}^{N} \mathbf{f}_{(n \mod (2L+1)q/p)}. \tag{10.17}$$

Since this FFT is of smaller size than the list $\mathbf{f}$, it is extended to the full size, $2L+1$, by periodically expanding $\hat{\mathbf{f}}^\dagger$.

There is a subtle difference between Eq. (10.12) and Eq. (10.16). The difference is that in Eq. (10.13) an infinite sum is taken over m and n. When this sum is truncated, as is done in Eq. (10.12), this yields a good approximation of a Dirac delta function only for a part of the domain of x and x' that is as wide as the region where the $b_i(x)$ peak is close to one (see Fig. 10.1(b)). Hence, this sets the coefficients $\mathbf{f}_i$ for large $|i|$ to zero. The main cause of differences between the Gabor-based Fourier transformation and the FFT-based Fourier transformation is the periodic continuation of functions at the edges of the domain in the region where the basis functions are close to zero.

To demonstrate the range over which the approximated Gabor transform is accurate, we apply the Fourier transformation operator on the modulated and shifted Gaussian pulse $\exp\{(x-x_0)^2 + jk_0x\}$. This pulse function is localized around (x_0, k_0) in spatial-spectral plane. We compare the discretized pulse function with a pulse function transformed to and from the spectral domain via Eq. (10.16). In Figure 10.3, the relative error is shown as a function of the location of the pulse (x_0, k_0) in the $x-k$-plane. Clearly, there is a four-digit accuracy over most of the domain, which corresponds to the accuracy up to which the dual window $\eta(x)$ was computed. For large k_0 and x_0 the accuracy is lower, because of the oversampling and the periodic continuation of functions. Therefore, we conclude that modulated Gaussian pulses can be accurately transformed back and forth from the spectral domain and since a discretization based on Gabor frames that consist of Gaussian pulses delivers accurate results, this method will be accurate as well.

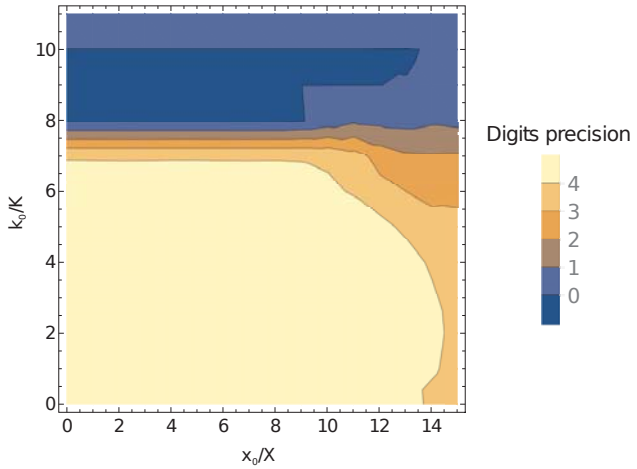

Figure 10.3: The relative error for the Fourier transform in a Gabor frame with $M = 15$ and $N = 13$ for a Gaussian pulse function located at a position in the spatial-spectral $x - k$-plane.

10.6 Application to a three-dimensional scattering problem

We exchanged the Gabor-frame discretization in [152] (Chapter 8) with the proposed list-based discretization. The Green function and the contrast function are not continuous enough to discretize by simply sampling them according to Eq. (10.7). Note that list-based representations of these functions are only calculated during the initialization of the algorithm, not during each iteration of the iterative solver. Therefore, the computation time of these lists is not very critical. We find a better approximation for these functions by taking the Gabor coefficients as they were calculated in [152] and transform them to the list-based representation through $\mathcal{B}^{-1}$.

The convergence of this method was tested on the first test case of [152], a dielectric cube embedded in a layered medium. Timing and accuracy results are displayed in Figure 10.4 for the two methods of Fourier transformation in Eq. (10.12) and in Eq. (10.16). Both employ the multiplication in Eq. (10.11). Clearly, the latter Fourier transformation requires much less computation time, while the results are very close in terms of accuracy. In the example, the difference between the results with different Fourier transformations is smaller than the error from the simulations itself, indicated by the circles in Figure 10.4. This implies that it is the discretization that governs the accuracy of the result, not the type of Fourier transformation that is being applied.

To show the applicability of this method to a larger problem, we compute the far field due to scattering from a finite grating consisting of 12 bars of relative permittivity $\varepsilon_r = 2.25$ in vacuum placed on a half space with $\varepsilon_r = 20.21 - 1.8j$ with a normal incident plane wave of unit amplitude as depicted in Figure 10.5(a). A Gabor frame with a window width of 1 micron was used that was truncated at $M = 7$ and $N = 16$ in both the x

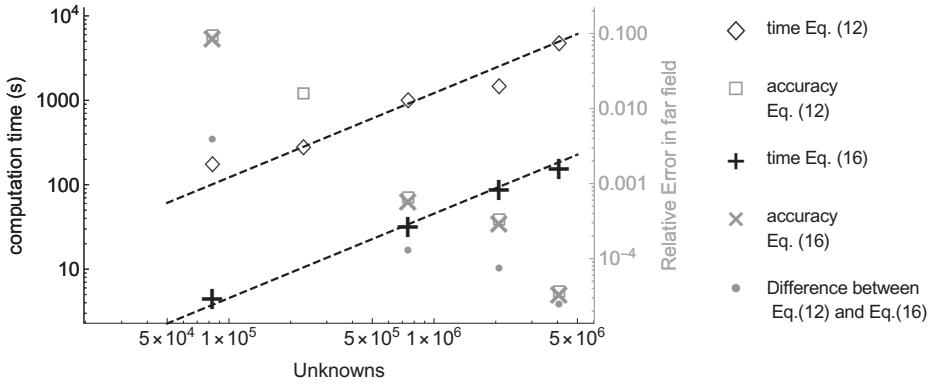

Figure 10.4: Scaling of the computation time and accuracy with the number of unknowns for different samplings in the xy plane, realized by truncating the Gabor frame at $N \in \{1, 4, 7, 10\}$. The relative error is computed against a JCMWave [134] reference calculation as described in [152] (Chapter 8). The dashed lines indicate linear scaling of computation time.

and y directions. In Fig. 10.5(b) and (c), the near and far field are depicted. In the far field, the relative $\mathcal{L}^2(\mathbb{R}^2)$ difference between results obtained with Eq. (10.16) and the algorithm in [152] was estimated at settings where the latter reached a 3×10^{-3} relative error. The computation time required with the present method was 105 minutes, using the Fourier transform in Eq. (10.12) it was 54 hours, and the time required by employing a Gabor-based multiplication [135] was estimated to be longer than 40 days based on the time required for a single multiplication. All computation times pertain to a single core of a 3.1 GHz Intel Xeon E5 2687W processor.

Since our complex path formulation works with a Gabor frame, and when Gabor-frame functions (modulated gaussians) can be accurately transformed with this newly developed Fourier-transform it will work. So this discretization should hold for discretizing the complete EM-scattering problem from dielectric objects. To show this, we show the electric field generated by a line source $J(x, z) = \delta(z)\Pi(x/100\mathtt{nm})$ observed at $z = 250$nm and $\lambda = 245$nm calculated using frame r' in Figure 10.6. The relative error in this plot is a bit larger than 10^{-3}. The reason is that a coarser spectral sampling is used compared to the direct Gabor coefficient multiplication. With this algorithm oversampling is not possible in the spectral domain, therefore the sampling in the spectral domain is much coarser than with the Gabor coefficient method. However, choosing a large oversampling has a larger penalty for the number of operations than simply choosing a larger spatial simulation domain. We verified that choosing the simulation domain twice as large does indeed lower the relative error below 10^{-3}.

We have benchmarked the calculation-time within Mathematica for a Fourier transform with Gabor frame r. The FFT-based method was 15 times faster with our Mathematica based algorithm, however, this is not a good indicator for performance in more optimized programming languages such as Fortran.

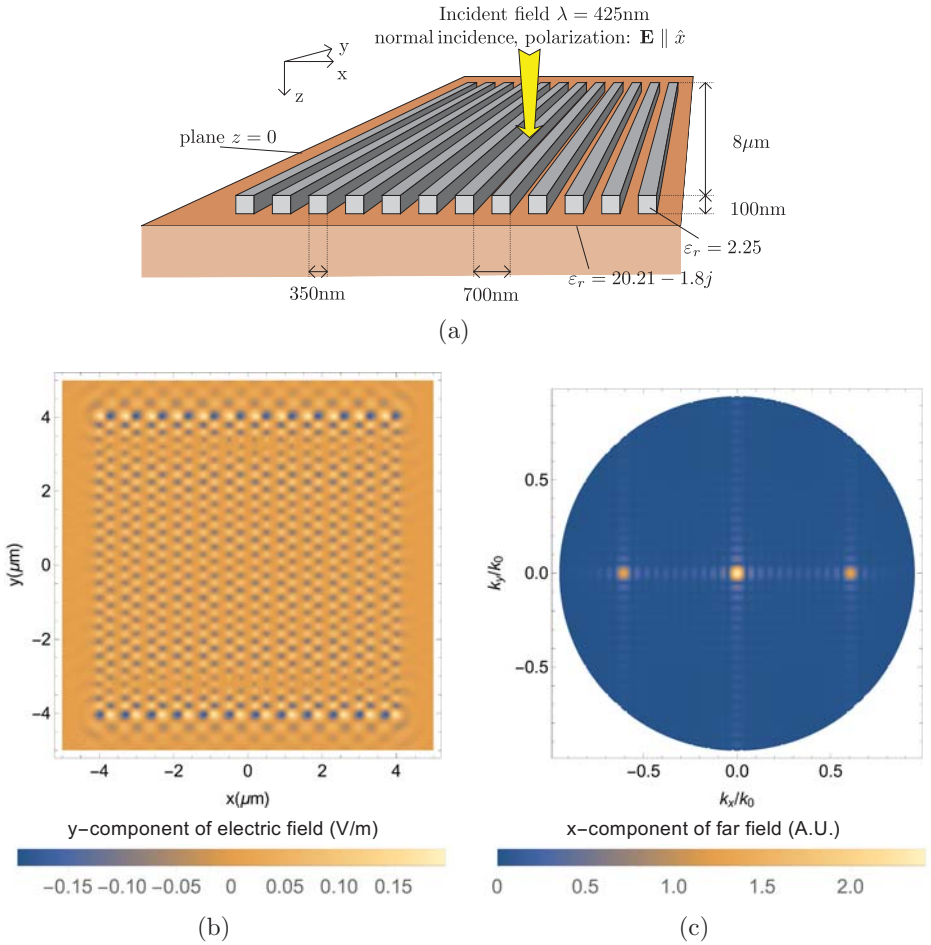

Figure 10.5: (a) A scattering setup of 12 dielectric lines on a dielectric halfspace. In (b) y-component of the electric field is plotted at $z = 20$ nm. In (c) the x-component of the far field is given that scatters back.

10.7 Conclusion

A point-wise multiplication and FFT-based Fourier transform operation were proposed based on a discretization by Gabor frames. An improvement to the algorithm in [152] is proposed that is at least 15 times faster for two represntative computational examples. Numerical evidence was shown that the approximation error is negligible compared to the discretization error.

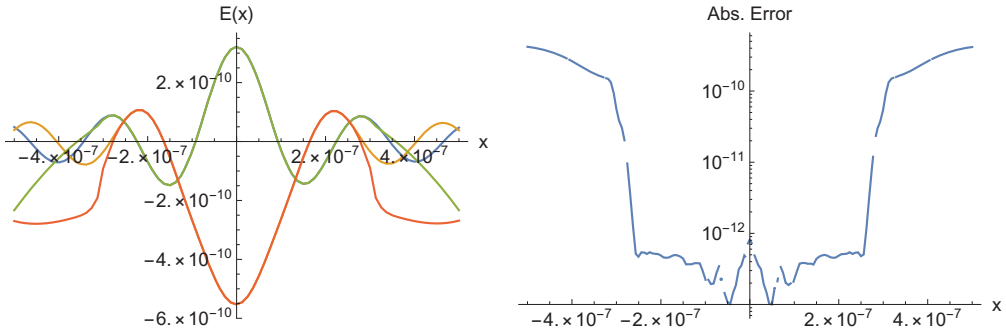

Figure 10.6: Performance of a caculation using frame r'. In (a) Blue and yellow are the real and imaginary part of numerical integration. Green and orange are the results generated with the presented method.

Addendum

To clarify the differences between the algorithm for 2D TM scattering in Chapter 7, the algorithm for 3D scattering in Chapter 8, and the algorithm presented in this chapter we present a schematic overview of the employed discretizations in Figure 10.7. The algorithm in Chapter 7 only employs the Gabor-frame discretization. In the description of this algorithm different names are used for the contrast function, $\mathcal{L}_\chi$ corresponds to $\mathcal{C}_\varepsilon$ and $\mathcal{M}_\chi$ corresponds to $\chi\mathcal{C}_\varepsilon$. The algorithm in Chapter 8 uses an equidistant sampled list for all operations except the Fourier transformation and the inverse Fourier transformation. A detour is taken via Gabor transforms to apply the Fourier transformation on Gabor coefficients. Initially, all components of the Green function, the reflection coefficients and contrast functions are all discretized via the Gabor frame. Afterwards, an inverse Gabor transform is employed to transform them to lists with equidistant sampling. The only thing that has changed in this chapter is that the Fourier transformation is carried out directly on the lists with equidistant samples with the method described in Section 10.5.2. The method in this chapter does not follow the detour via the Gabor transformations, it employs the direct Fourier transformation that is described in this chapter.

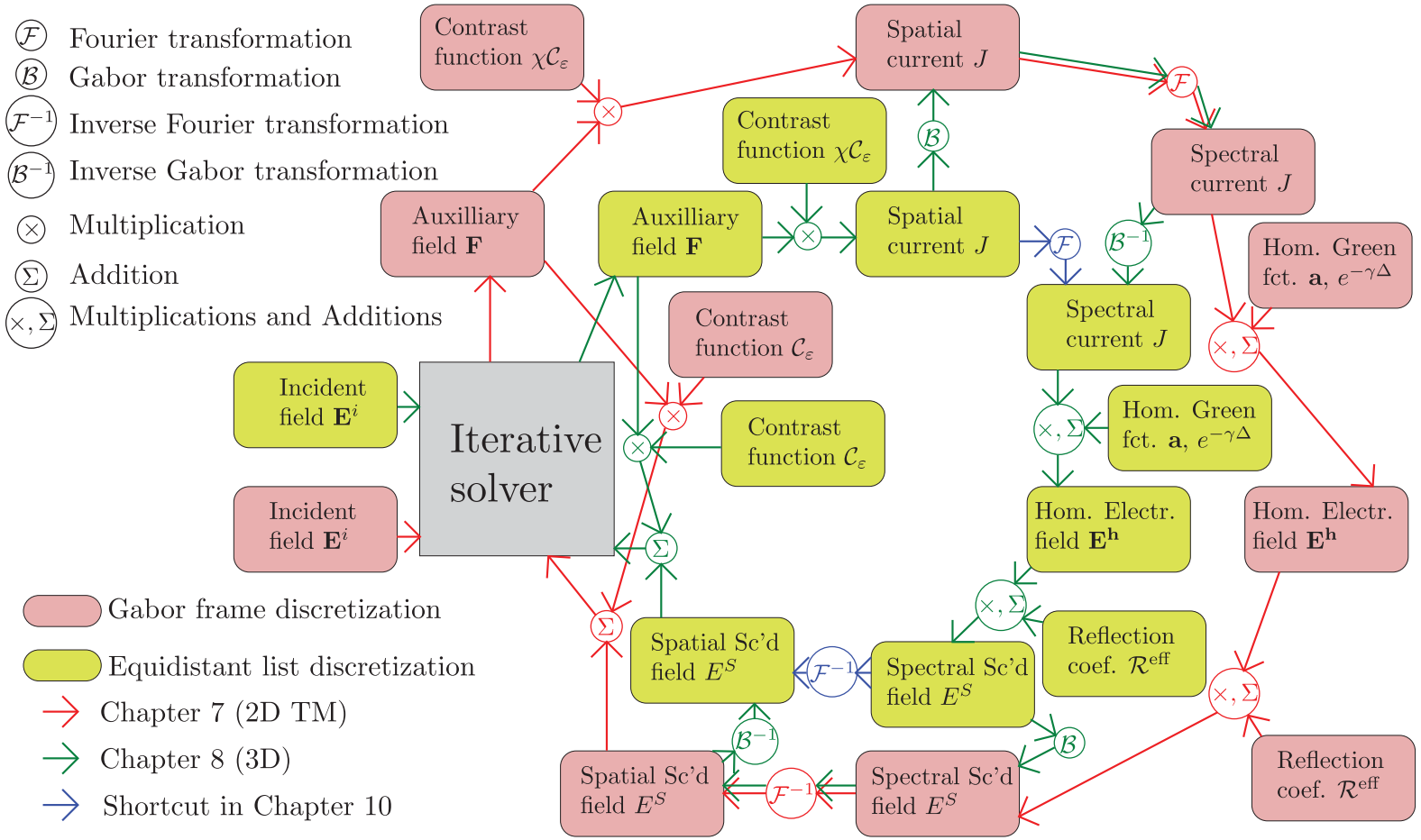

Figure 10.7: A schematic description of the steps required in the algorithms of Chapter 7, Chapter 8 and this chapter. By Sc'd field, the scattered field is meant.