Research Repository UCD

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	Fast linear canonical transforms
Authors(s)	Healy, John J.; Sheridan, John T.
Publication date	2010-01-01
Publication information	Journal of the Optical Society of America A, 27 (1): 21-30
Publisher	Optical Society of America
Link to online version	http://dx.doi.org/10.1364/JOSAA.27.000021
Item record/more information	http://hdl.handle.net/10197/3296
Publisher's statement	This paper was published in Journal of the Optical Society of America. A, Optics and image science and is made available as an electronic reprint with the permission of OSA. The paper can be found at the following URL on the OSA website: http://www.opticsinfobase.org/josaa/abstract.cfm?uri=josaa-27-1-21. Systematic or multiple reproduction or distribution to multiple locations via electronic or other means is prohibited and is subject to penalties under law.
Publisher's version (DOI)	10.1364/JOSAA.27.000021

Downloaded 2022-03-05T03:23:43Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information, please see the item record link above.

Fast linear canonical transforms

John J. Healy[1,2,3,4] and John T. Sheridan[1,2,3,*]

[1]*UCD Communications and Optoelectronic Research Centre, College of Engineering, Mathematical and Physical Sciences, University College Dublin, Belfield, Dublin 4, Ireland*
[2]*SFI Strategic Research Cluster in Solar Energy Conversion*
[3]*School of Electrical, Electronic and Mechanical Engineering, College of Engineering, Mathematical and Physical Sciences, University College Dublin, Belfield, Dublin 4, Ireland*
[4]*Complex Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Ireland*
**Corresponding author: john.sheridan@ucd.ie*

Received August 25, 2009; accepted November 2, 2009;
posted November 4, 2009 (Doc. ID 116174); published December 3, 2009

The linear canonical transform provides a mathematical model of paraxial propagation though quadratic phase systems. We review the literature on numerical approximation of this transform, including discretization, sampling, and fast algorithms, and identify key results. We then propose a frequency-division fast linear canonical transform algorithm comparable to the Sande–Tukey fast Fourier transform. Results calculated with an implementation of this algorithm are presented and compared with the corresponding analytic functions.
© 2010 Optical Society of America
OCIS codes: 070.4560, 080.2730, 100.2000, 200.2610, 200.3050, 200.4560, 200.4740.

1. INTRODUCTION

The Fresnel transform provides a well-known mathematical model of paraxial free space propagation [1]. Discrete approximations to it are used for reconstruction of digital holograms [2]. The linear canonical transform (LCT) can be used to model systems composed of lenses and free space and other quadratic phase systems, e.g., graded-index media [3,4]. Named for its area-preserving, linear coordinate transforming effect on the phase space description of a wave field (e.g., its Wigner–Ville distribution function (WDF) [5,6]), the LCT is a parameterized linear integral transform. The parameters of the transform are related to the ABCD or Collins matrix characterization of the system used in ray-tracing calculations [1]. The LCT's origins are in quantum mechanics; a brief overview may be found in [7]. Although it was considered for such use much earlier, e.g., [8], the LCT's application in scalar diffraction theory has gained further recognition since the fractional Fourier transform (FRT) was first introduced in optics. The LCT has also been discussed as a generalization of the FRT [3,9]. A review of the history of the LCT in optics is presented in [4], which also discusses the transform's properties.

There are a number of attractive facets to an LCT-based discussion of optics: (i) it provides links to geometric optics via the ray-tracing matrix that characterizes the optical systems in both; (ii) it has the potential to draw mathematical tools from the vast literature on Fourier analysis; and (iii) it allows simple relationships to be drawn between phase space optics and time–frequency representations. Aside from the Fresnel transform and the FRT, the LCT's special cases also include the Fourier transform (FT), the effect of a thin lens (chirp multiplication), and magnification (scaling). We restrict our discussion here to lossless systems and therefore to LCTs with real parameters. However, LCTs with complex parameters are also of interest, e.g., the Laplace transform and the Gauss–Weierstrass transform [7].

The continuous LCT has been used in a number of applications, e.g., in optical systems to measure tilt and translation [10–12], to provide additional keys in double-random phase encoding optical encryption schemes [13,14], in analysis of speckle size [15], and in noninterferometric phase extraction schemes [16,17]. However, there are many situations where it is desirable to numerically approximate the transform. A comparison may be made with the utility of the ubiquitous fast Fourier transform (FFT) algorithms for numerically approximating the FT. Fast, accurate, and simple numerical tools for approximating the LCT are particularly useful in situations where discrete data from a digital camera must be processed, e.g., in digital holography. Such a discretization of the transform appears to have been first undertaken by Pei and Ding [18], and while their formulation remains the accepted one in praxis [19], it presents certain practical inconveniences for users [20], particularly in relation to sampling issues [21]. A new definition of the discrete LCT (DLCT) has recently been proposed [22,23], along with a sampling methodology that appears to address the issues raised in [20,21]. We will review this methodology in a subsequent section. We have chosen to use the older definition of the DLCT [18] in this paper, but as we will show in Section 2, there is little practical difference between it and that of [22,23] other than the reconstruction filter used.

The focus of this paper is efficient, fast algorithms for

 © 2010 Optical Society of America

evaluating the DLCT. These are to the DLCT what the FFT is to the discrete FT (DFT). Numerical approximation of the transformation is desirable for the development of simulation tools. Considering the ubiquitous use of the FFT [the generic term for numerical algorithms for calculating the DFT with $O(N \log N)$ complexity], analogous fast algorithms for the LCT may prove to be valuable outside of their obvious application in simulating optical systems. The discrete calculation can be thought of as a matrix multiplication, $f_M = L_M f$, where f is a $1 \times N$ vector consisting of the samples of the input wave field, f_M is the vector of samples of the output field, and L_M is the $N \times N$ discrete transform matrix for ABCD matrix M. The direct evaluation of this multiplication is of $O(N^2)$ complexity. Fast algorithms exploit redundancies in the matrix to iteratively break the multiplication into multiple smaller calculations, reducing the overall complexity. The first direct fast LCT (FLCT), a radix-4 mixed time- and frequency-division algorithm, was proposed in [19]. It iteratively decomposed the matrix multiplication into four smaller ones. A second, alternative approach to numerical approximation of the LCT is based on decomposing the transform or, equivalently, the discrete transform matrix, into a series of special cases for which fast algorithms are known, e.g., the DFT and the discrete FRT. This avenue of research has thus far culminated in [24]. It is not yet clear which of these two types of algorithm, if either, is most useful. However, decompositions that use the FT do benefit from the availability of highly optimized FFT implementations.

In overview, this paper is organized as follows. In Section 2, we define the LCT and DLCT and review relevant properties of the transform and sampling theory. We compare Pei and Ding's DLCT [18] and the associated sampling methodology [21,25] with those presented in [22,23]. In Section 3, we review algorithms for calculating the DLCT, including direct and decomposition-based. In Section 4, we derive a decimation-in-frequency FLCT algorithm, comparable to the Sande–Tukey frequency-division FFT algorithm. In Section 5, we present the results of some calculations performed using this algorithm, including comparisons with corresponding analytical results. Finally, we present our conclusions in Section 6.

2. LCT: DEFINITION, PROPERTIES, SAMPLING, AND DISCRETIZATION

Given a system of lenses and free space, we may characterize the system in the paraxial approximation using an ABCD matrix, or ray-transfer matrix [1]. In ray tracing, this matrix relates the position and angle of rays at the input and output of a system. The matrix is symplectic. The ABCD matrix for free space (and hence the Fresnel transform) is $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & \lambda z \\ 0 & 1 \end{smallmatrix}\right)$, where λ is the wavelength of the propagating quasi-monochromatic light and z is the propagation distance. The ABCD matrix for a thin lens of focal length f is $\left(\begin{smallmatrix} 1 & 0 \\ -1/\lambda f & 1 \end{smallmatrix}\right)$. If a system consists of several sequential sections for which the ABCD matrices are known, the total system matrix is given by the product of the ABCD matrices of the subsections. The inverse problem, i.e., decomposition of a system's ABCD matrix, is relevent for algorithm analysis and design [24,25] and system design [26]. While the two matrices above are of particular physical significance for optical system design, designers of numerical algorithms may prefer to employ different "building blocks." Thus while chirp multiplication remains an important numerical operation, the FT, scaling, and even the FRT are more important for algorithm design than free space propagation, because mature numerical algorithms for calculating them make decompositions using those transforms faster than the alternatives.

We will now define the 1D LCT for a given optical system. Generalization to the 2D case is usually straightforward since the transform is separable for orthogonal systems. A different definition must be used for nonorthogonal systems, which lack symmetry about the optical axis e.g., because they contain pyramidal mirrors or cylindrical lenses (except where the axes of the lens line up with our coordinate system) [27,28]. However, we will not consider such systems in this paper, and we note that all the existing algorithms for calculating the LCT are unsuitable for performing simulations of such systems. Consider a system characterized by some ABCD matrix, M. Let the wave field at the input plane be described by $f(x)$ and at the output plane by $f_M(y)$. These two functions are related as follows:

$$f_M(y) = L_M\{f(x)\}(y) = \begin{cases} \sqrt{\dfrac{1}{2\pi B}} \exp\left(\dfrac{-j\pi}{4}\right) \displaystyle\int_{-\infty}^{\infty} f(x) \exp\left\{\dfrac{j}{2B}(Ax^2 - 2xy + Dy^2)\right\} dx, & B \neq 0 \\ \dfrac{1}{\sqrt{A}} \exp\left(\dfrac{j}{2}\dfrac{C}{A} y^2\right) f\left(\dfrac{y}{A}\right) & B = 0 \end{cases}, \tag{1}$$

where $f_M(y)$ is the LCT of $f(x)$. The transform is affine, additive, unitary, and invertible. The inverse is simply the LCT with parameters found by inverting the original ABCD matrix of the forward transform. Throughout the remainder of this paper, we will consider only systems with $B \neq 0$ except where specifically noted. When $B = 0$, the transform is simply a chirp and a scaling. This is trivial to implement numerically as the corresponding discrete transform matrix is diagonal, and the operation therefore has $O(N)$ complexity.

A. Sampling

Key to a successful numerical approximation of the LCT is a clear picture of the consequences of sampling a signal and any LCT of that signal. The literature on this subject remains active. However, an overview of its development to date is enlightening. Analyses based on the WDF-based phase space diagrams (PSDs) popularized by Lohmann are especially illuminating. Most of these make assumptions about the signal and its bandwidth, but as Onural has pointed out [29], it is important to bear in mind that assumptions about bandwidth may not necessarily be the most efficient or convenient assumptions to make about scalar wave fields. While a theorem analogous to the classic Shannon sampling theorem has been known to be applicable to the LCT for almost a decade, its relevance to the numerical approximation of the LCT has only recently become clear.

Shannon sampling assumes that a signal's FT is nonzero only over a finite range of frequencies. The width of this support is then directly related to a sampling rate for the signal, and a reconstruction filter is specified that allows recovery of the continuous signal from its samples taken at not less than that rate, which is known as the Nyquist rate. As the FT is a special case of the LCT, one may consider what the analogous sampling rate and reconstruction process are for a signal that has compact support in some LCT domain. Gori established this result for the Fresnel transform in 1981 [30]. We will not discuss this theorem or its proof except to note that it involved rewriting the transform in terms of a FT and applying Shannon's theorem. In the same paper, Gori also proved that a signal with finite bandwidth could not have finite support in any Fresnel domain. The equivalent of this latter result for the LCT is more complex [31], depending on the specific ABCD parameter values. Various authors have established the sampling theorem for signals that are compact in some fractional Fourier domain, of which Xia seems to have been first [32]. To the authors' knowledge, Ding [33] was first to propose the equivalent sampling theorem for the LCT. While no specific name is established for this theorem, we will refer to it as Ding's theorem throughout the remainder of this paper. Ding shows that if a particular LCT of a signal is zero outside some range $-L/2<x<L/2$, then the signal may be sampled at regular intervals of $T\leqslant 2\pi B/L$. His proof is similar to Gori's proof for the Fresnel case, though he does not reference [30]. Both Gori's and Xia's theorems are special cases of Ding's. This LCT sampling theorem was later independently derived by Stern [34], whose proof differs from Ding's. Stern uses the Poisson formula for the comb function to derive the LCT of a sampled signal, which he shows consists of modulated, periodic copies of the LCT of the original, continuous signal, a property referred to as "chirp periodicity." If the LCT of the continuous signal has compact support at least approximately over a finite range, and if the sampling rate used is sufficient to prevent the replicas from overlapping, it is possible to extract one copy from among the replicas using an appropriate filter. Deng *et al.* [35] published a derivation of the same theorem almost simultaneously with that of Stern (the two papers being submitted less than three weeks apart). Their approach was to derive the convolution and multiplication theorems for the LCT and then calculate the LCT of the product of an arbitrary function and a train of delta functions. Like Stern, they showed that the LCT of a sampled signal consists of modulated, periodic copies of the LCT of the original, continuous signal, thus arriving at the same conclusion. While Stern approached reconstruction as simply multiplying by a rect function in the appropriate domain, Deng *et al.* explicitly derived the time-domain reconstruction formula. Recently, Li *et al.* [36] published a proof of Ding's theorem similar to that in [33].

The reconstruction formula, as defined by Stern, is given by

$$f(x)=TL_{M^{-1}}\left\{\mathrm{rect}\left(\frac{y}{L}\right)L_M\{f(nT)\}(y)\right\}(x), \tag{2}$$

where

$$\mathrm{rect}(x)=\begin{cases}1 & |x|<0.5\\ 0.5 & |x|=0.5\\ 0 & |x|>0.5\end{cases}. \tag{3}$$

We note that Zhao *et al.* recently considered rate conversion of a signal sampled using this theorem [37].

We will shortly discuss a proposal for practical use of Ding's LCT sampling theorem. First, however, we consider another proposed solution to the following important, practical question: Given a particular wave field that is to propagate through a system characterized by some ABCD matrix, how should we sample that wave field and the field appearing at the system's output? This question was addressed in detail in [25] by examining the LCT's effect on the PSD of a signal. If a signal, $f(x)$, has a WDF, $W\{f(x)\}(x,k_x)$, and an LCT, $f_M(y)=L_M\{f(x)\}(y)$, then the WDF of $f_M(y)$ is [4]

$$W\{f_M(y)\}(y,k_y)=W\{f(x)\}(Dx-Bk_x,-Cx+Ak_x). \tag{4}$$

This interpretation of the LCT has led to proposals for the transform's use in numerical filter design [38] and proves useful in the discussion of sampling processes. In [25], it was assumed that $f(x)$ had some known spatial extent and bandwidth, where these were defined as the widths in the spatial and frequency domains within which all but an arbitrarily small portion of the total signal power was contained. This defined what Lohmann called the "space–bandwidth product shape" [39], or what more recent literature refers to as the signal's PSD. This appears as a rectangle in phase space subtending a portion of the WDF in which most of the signal's power is localized. Total localization can be shown to be impossible mathematically [31]. Equation (4) describes the effect an LCT has on a signal's WDF, transforming the initially assumed rectangle into a parallelogram. Lohmann used this to determine the effect of LCTs on a PSD. In [25], the resulting PSD parallelogram is used to define an extent and bandwidth for the transformed signal, and it is also applied to tracking the evolution of the PSD as the signal passes through an optical system or as it is processed in an algorithm based on decomposition of the LCT into simpler transforms. However, the sampling criterion established in [25] is not sufficient. A second criterion was recently

added to allow for the effect of discretization of both the input and the output domains [21].

We now consider a recently proposed sampling methodology [22,23] that makes use of Ding's LCT sampling theorem rather than Shannon sampling. There are benefits to this approach in terms of the number of samples typically used to describe the signal. The only drawback to the approach is that the reconstruction filter is more complex than for the methodology considered in the preceding paragraph. As before, it is assumed that a wave field has some known spatial extent, but the second assumption here is not finite bandwidth but finite support in the LCT domain in question. The LCT of a signal so sampled consists of modulated replicas of the LCT of the continuous signal. This signal is then sampled, once again using Ding's theorem. The sampling rate for the first sampling operation depends on the extent of the LCT of the signal, while the sampling rate for the second depends on the extent of the signal itself. This greatly simplifies the sampling requirements, as the input signal's extent is usually well known, leaving just one parameter, the output width, to adjust. Unlike the methodology of [25], no additional calculations are required to establish additional parameters. We will discuss this in greater detail in Subsection 2.B, where we discuss the DLCT.

B. Defining the DLCT

In this section, we first introduce the definition of the DLCT used in this paper. We also review the DLCT proposed in [22,23], contrasting and comparing it with the one used here. It should be noted that regardless of the advantages of the two different DLCT definitions, the direct algorithms developed here and in [19] are compatible with both.

In this paper, we use the same definition for the DLCT as was used in [19], which is equivalent to the DLCT defined by Pei and Ding in [18]. For a discrete function f, consisting of N samples, at a sampling rate T in the input domain and $T_y = \beta T$ in the output domain, the DLCT is as follows,

$$D\Theta_{\alpha\beta\gamma}^{T,N}\{g(nT)\}(mT_y) = \sqrt{\frac{\beta}{2\pi}}\exp\left(\frac{-j\pi}{4}\right)\sum_{n=-N/2}^{(N/2)-1} g(nT)W_N^{n,m}, \tag{5a}$$

where

$$W_N^{n,m} = \exp\{j\pi[\alpha(nT)^2 - 2nm/N + \gamma(m/NT\beta)^2]\}. \tag{5b}$$

In this equation, we have taken the range of the input field to be zero centred in accordance with the convention for optical systems with a principal axis. However, we will modify this using the appropriate shift theorem so that the range of n is from 0 to $N-1$ in order to simplify the derivation of the algorithm. This convention is typical for FFT algorithms, as the signal processing community most commonly deals with temporally varying signals. This definition of the DLCT uses the parameters α, β, and γ instead of the ABCD parameters, but these are trivially equivalent [4]. In [20], it is described how to set the various output parameters (range, bandwidth, etc.) by altering the sampling rate and/or zero padding at the input.

As noted, an alternative definition of the DLCT was first proposed in [22] and then independently in [23]. In [23], the merits of this alternative definition are discussed. We now contrast and compare this new DLCT with that previously defined by Pei and Ding [18], the latter definition being the definition of choice thus far for direct FLCT algorithms. The literature to date on decomposition-based algorithms, e.g., [24], has not explicitly used a DLCT, but we believe the formulas implicit in those algorithms are essentially of the form of the DLCT of [18]. Accordingly, this discussion may also have relevance for those algorithms. One consequence of the sampling methodology of [22,23] is that the output samples must be recovered in a fashion consistent with the sampling theorem used. If the signal is be interpolated, LCT-specific rate conversion methods [37] should be used. However, we note that it is also possible to make such sampling assumptions for the DLCT of Pei and Ding. We claim that this sampling methodology [22,23] uses far fewer parameters, i.e., variables that must be chosen by the user, than that of [21,25] and furthermore should typically result in a significantly lower number of samples. For these reasons, we believe it may become the standard method used in this field. We now justify both of these claims.

As we briefly mentioned in the section on sampling (Subsection 2.A), the most important features of both DLCT definitions are illustrated by the PSD analysis shown in Fig. 1 and Fig. 2. Figure 1 illustrates the sampling process of [25], where we begin with the assumption that the signal approximately has some finite bandwidth. Together with the width of the input, this defines the input rectangle in Fig. 1(a). By way of contrast, Fig. 2,

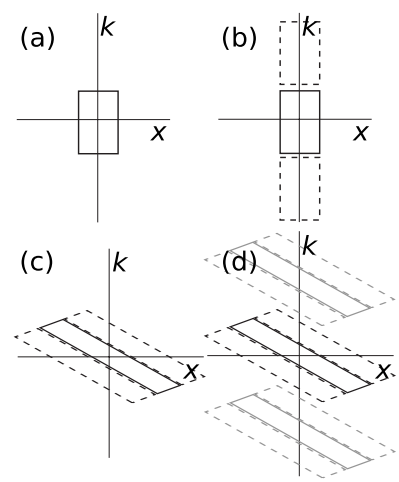

Fig. 1. Sampling according to [25]. (a) The signal is assumed to be bounded in space, x, and in the Fourier domain, k. (b) The signal is sampled at a rate not less than the Nyquist rate, making it periodic (but not overlapping) along the frequency axis (dashed lines indicate replicas). For convenience, only the two closest replicas are shown. (c) An LCT operation is performed on the discrete signal. (d) The transformed signal is sampled (gray lines indicate replicas created by this process). Note the irregularity of the replicas, which may complicate recovery of the continuous signal from the samples or even introduce aliasing effects for some choices of sampling rates as described in [21]. In this case, it is clear that if we include more replica terms from the first sampling operation, there will be some overlap with the zeroth-order term.

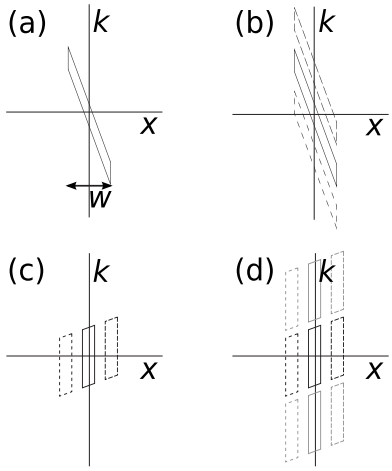

Fig. 2. Sampling according to [22,23]. (a) The signal is assumed to be bounded in space and in the output linear canonical domain. (b) The signal is sampled at a rate not less than that given by Ding's LCT sampling theorem, making it periodic along the frequency axis (dashed lines indicate replicas). For convenience, only the two closest replicas are shown. (c) An LCT operation is performed on the discrete signal. (d) The transformed signal is sampled (gray lines indicate replicas created by this process). Note the regularity of the replicas—a consequence of the bounds chosen for the original signal. These diagrams have been scaled to provide clearer illustration, but this does not alter the argument.

which illustrates the process of [22,23], begins with an assumption of finite width of the LCT, resulting in the parallelogram of Fig. 2(a). In both processes, the signal is sampled [parts (b)], the LCT of this sampled signal is obtained [parts (c)], and the resulting transformed signal is sampled [parts (d)]. We note that the process described in Fig. 2 does not occur in the same order as described in [23]. In that paper, it is assumed that the signal to be sampled is chirp periodic, while our discussion, somewhat like that of [22], considers all periodicity to arise from the two sampling operations. However, we note that the descriptions are completely equivalent. The sampling process of [22,23] and Fig. 2 is defined in such a way that the "pattern" of replicas is automatically orderly; i.e., if both the sampling operations are performed at exactly the generalized Nyquist rates given by Ding's theorem, the replicas will completely cover the phase space plane without any overlap. This reduces the number of samples used to represent the signal, resulting in consequent benefits in terms of computation time.

The area of the parallelogram chosen using the Stern methodology [22,23], like that of the rectangle chosen using [25], i.e., the *space–canonical width product* and the *space–bandwidth product* (SBP), both indicate the information content of the signal. These areas may be similar for a given signal or vary considerably, depending on the structure of the signal's WDF. Without making further assumptions about the structure of the WDF, we must conclude that whichever measure is greater is random. If the output SBP is greater than that of the input (and it must be not less than it if we assume a rectangular input PSD), the larger number of samples must be used to perform the computation. In that case, the methodology of [25] will typically use more samples than that of [22,23]. We note that while there may be signals for which this is not the case, the additional burden imposed by [21] means that [22] is likely to remain more efficient than [25] even in these special cases.

There is also an important practical advantage to [22,23] in terms of ease of use. Given a wave field that is to be sampled and then transformed according to [21,25], we must first estimate that signal's width and bandwidth. These are multiplied to determine the number of samples needed to represent the signal, i.e., the input signal's SBP. We must then perform a pair of matrix multiplications to determine the corresponding parameters for the output wave field. The larger of the two numbers of samples determines the number of samples used in the calculation. Careful consideration must then be made of how to best choose parameters such as the output extent [20]. In contrast, given a wave field that is to be sampled and transformed according to [22,23], we need only estimate its input width and the width of the output wave field. All other considerations are then automatically taken care of. This is similar to the operation of the DFT, where the width of the input signal defines the sampling rate of the output and vice versa. The simplicity of [22,23] and its consistency with the operation of the DFT offers a clear advantage over [21,25].

The elements of the DLCT matrix in [23] are independent of the sampling rates in the input and output planes of the optical system, i.e., the LCT domains. They depend on the product of these sampling rates, which product is proportional to the number of samples used. We note that this is achieved by scaling the signal such that it has the same width as its transform. It is interesting to compare this procedure with the prescaling operation used in Koç *et al.*'s decomposition-based LCT algorithms [24] and the Lohmann–Rhodes light tube model for Fresnel transforms [40], both of which similarly scale the input. In the particular case of the FT, the fact that the matrix does not depend on the sampling rates is significant in one regard: it facilitates optimization of FFT algorithms. This is because we may optimize code or hardware for a particular number of samples, e.g., 2048 [41]. These algorithms typically contain many multiplications of the data by constants, which can be hard coded (or wired). In the case of the LCT, the elements of the transform matrix depend on the transform parameters and so cannot be similarly hard coded/wired. Therefore, removing the dependence of the discrete transform matrix on the sampling rates, as is done in [23], is of significant practical value only if one wishes to calculate very many transforms for the same optical system parameters and sampling rates.

In conclusion, the sampling methodology proposed in [22] is significantly simpler and more efficient than methodologies previously described in the literature. [23] modifies this DLCT definition [22] to allow the elements of the discrete transform matrix to be independent of the sampling rates at the input and output. While this definition is more consistent with the DFT, it is not in general as advantageous as in that special case.

3. FLCT ALGORITHMS

The first direct fast algorithm for the DLCT was presented in 2005 [19]. It involved a mixture of frequency

and time division. As far as we are aware, the special case of this algorithm for the FT has not appeared elsewhere and may itself be of interest. Another paper, [25], published simultaneously, reviewed many of the existing algorithms for calculating the FRT, the Fresnel transform, and the LCT and showed that they consist of a variety of decompositions of the ABCD matrix. It was also demonstrated that tracking the signal PSD, and therefore the sampling requirements, as the signal passed through the transform modules (corresponding to the decomposed matrices) explained why many of these algorithms were previously thought to be nonunitary or even useful only over certain ranges of their parameters. Applying such a decomposition technique, an implementation [42] based on the decomposition of the LCT into an FRT, magnification, and chirp multiplication has been proposed. The fast FRT algorithm used was one of two presented in [43], one of which is itself based on a decomposition of the FRT into a chirp multiplication followed by a chirp convolution followed by another chirp multiplication. The other FRT algorithm in [43] uses an FFT-based convolution and a chirpmodulation. Most recently, Koç *et al.* [24] exhaustively analyzed decomposition-based algorithms and proposed that only two be used, with the decision between them depending on the ABCD parameter values. As noted, that paper also proposed the use of a prescaling method that permitted a determination of the sampling rates for these algorithms.

4. FREQUENCY DIVISION FLCT ALGORITHM

In this section, we present a fast algorithm with more flexibility in the number of samples with which it can operate than that discussed in [19]. We derive a general radix decimation-in-frequency fast algorithm to calculate the DLCT in the style of the Sande–Tukey FFT algorithm [44]. Later, in Section 5, we present results produced by an implementation of this algorithm for radix 2, which allows vector lengths of size 2^n to be transformed, where n is any integer. As the radix-2 algorithm, this implementation is no more flexible than that presented in [19], but because the algorithm is derived for an arbitrary radix, it is possible to implement it for any prime radix p, i.e., p^n samples. Concatenations of such algorithms are then sufficient to calculate the DLCT of almost any vector length (though we note that in practice, some nonprime blocks are also used in FFT implementations, as discussed in [45]). For example, additional implementations of radix 3 and radix 5 would allow the decomposition of any DLCT of a vector whose length, N, had prime factors 2, 3, and 5. Thus, for example, 1440 samples could be implemented as one radix-5 decimation, followed by two uses of a radix-3 implementation and five stages using the radix-2 algorithm. Using just a handful of different radices permits significantly more flexibility in choosing the number of samples than is the case when using a single-radix algorithm. This can provide considerable computational savings.

It is more difficult to deal with a prime number of samples. For the case of the FT, several $O(N \log N)$ algorithms are known for N prime, the earliest of which is due to Rader [46]. However, as yet, no prime-radix algorithm is known for the LCT. This remains a subject of interest, at least from the theoretical perspective of computer science.

We begin by defining the DLCT, in Eq. (6). This differs from the definition in Eq. (5a) and (5b) only in the range of the sum (which we now take from 0 to $N-1$) and in that we have, for brevity, replaced the operator notation and the sampled input by discrete input and output sequences. The former change simplifies the analysis somewhat, and both produce forms closer to the conventional notation used in signal processing for the FFT. Furthermore, the LCT-shifting theorems allow us to efficiently convert the result in Eq. (6) to a zero centered one, in an $O(N)$ calculation:

$$L[m] = \sqrt{\frac{\beta}{2\pi}} \exp\left(\frac{-j\pi}{4}\right) \sum_{n=0}^{N-1} f[n] W_N^{n,m}. \tag{6}$$

For simplicity, we neglect the $\sqrt{\beta/2\pi}\exp(-j\pi/4)$ term in the derivation below, as it does not affect the derivation. We assume that the number of samples, N, is nonprime; i.e., there exist integers r and s such that $N=rs$. We refer to r as the radix of the algorithm. We define sequences $f_0, f_1 \ldots f_i \ldots f_{r-1}$ of index $0 \leqslant n < N/r$ such that

$$f_i[n] = f[n + iN/r]. \tag{7}$$

This allows us to rewrite the expression for $L[m]$ in terms of r smaller sums

$$L[m] = \sum_{n=0}^{(N/r)-1} \sum_{i=1}^{r} f_i[n] W_N^{n+iN/r,m}. \tag{8}$$

We note that

$$W_N^{n+iN/r,m} / W_N^{n,m} = \mu_1 \mu_2,$$

where

$$\mu_1 = \exp[-j2\pi i m/r],$$

and

$$\mu_2 = \exp\{j\pi\alpha(N/r) i T^2 [2n + iN/r]\}. \tag{9}$$

Substituting back into Eq. (8), we get

$$L[m] = \sum_{n=0}^{(N/r)-1} \sum_{i=0}^{r-1} \mu_1 \mu_2 f_i[n] W_N^{n,m}. \tag{10}$$

We note that μ_1 is a function of m. We define the sequences $L_0, L_1, L_2 \ldots L_c \ldots L_{r-1}$, of index $0 \leqslant m < N/r$ such that

$$L_c[m] = L[rm + c]. \tag{11}$$

We also note the identity,

$$W_N^{n,m} = W_{pN}^{n,pm}, \tag{12}$$

which holds true for any constant p. This identity is used in many FFT algorithms, as is the equivalent scaling of n and N. However, the latter scaling is not true for the DLCT kernel, unless we also scale T, the input sampling period. This has implications for a decimation-in-time FLCT algorithm, which therefore can be implemented

only by altering the LCT parameters at each stage to counteract the magnification implicit in scaling T. This occurs because, unlike the FT, the shape of an LCT varies with magnification of the input. Using Eq. (12), we obtain

$$L[m] = \sum_{n=0}^{(N/r)-1} \sum_{i=0}^{r-1} \mu_1 \mu_2 f_i[n] W_{N/r}^{n,m/r}. \tag{13}$$

To proceed, we first define

$$k = (m - c)/r, \tag{14}$$

and we note that

$$\mu_3 \mu_4 = W_N^{n,k+c/r} / W_N^{n,k}, \tag{15}$$

where μ_3 and μ_4 are defined as follows:

$$\mu_3 = \exp(-j2\pi nc/rN),$$

$$\mu_4 = \exp[j\pi\gamma c(2kr + c)/(NT\beta r)^2]. \tag{16}$$

Equation (13) is rewritten as

$$L_c[k] = \mu_4 \sum_{n=0}^{(N/r)-1} \sum_{i=0}^{r-1} \mu_1 \mu_2 \mu_3 f_i[n] W_{N/r}^{n,k}. \tag{17}$$

While μ_1 is a function of m, and therefore of k, we note that for a given c, it is constant. Specifically, it can be written as

$$\mu_1 = \exp[-j2\pi c/r]. \tag{18}$$

For the purposes of clarity, we define a new sequence as follows:

$$f_c'[n] = \sum_{i=0}^{r-1} f_i[n] \mu_1 \mu_2 \mu_3. \tag{19}$$

Substituting into Eq. (18), we get

$$L_c[m] = \mu_4 \sum_{n=0}^{(N/r)-1} f_c'[n] W_{N/r}^{n,k}. \tag{20}$$

Thus, we can calculate the DLCT of a length N series of samples as the weighted sum of r-size N/r sequences. Performing this recursively allows us to calculate the overall DLCT in $O(N \log N)$ time.

We believe, in agreement with [22], that it is possible to evaluate Stern's DLCT using this algorithm with only minor modification.

5. NUMERICAL RESULTS

In this section, we present some illustrative results calculated using an implementation in C [47] of the radix-2 case of the algorithm considered in Section 4. We compare the output of this program with a MATLAB [48] implementation of the naïve $O(N^2)$ implementation of the DLCT to determine whether any effects are introduced by the fast algorithm. We also compare the numerical results to the corresponding analytic expression to confirm that the algorithm satisfactorily approximates the continuous transform.

First, we derive an analytical result with which to compare the output of the algorithm. An on-axis plane wave impinges upon a rectangular aperture of arbitrarily chosen width 9.96 cm (chosen to give around 102 samples at 1024 samples/m) before propagating through a lens system with an ABCD matrix, which yields LCT parameters $\alpha = 636.619249$, $\beta = 636.620034$, and $\gamma = 636.619249$. This can be implemented using a single lens, but implementations are outside the scope of this paper. We also assume $\lambda = 1$ (the effect of the wavelength is simply to alter the parameters, so this is not particularly significant). Mathematica [49] produces the following analytic solution:

$$L_M(y) = k \exp(-jly^2)\{\mathrm{erfi}[(1+j)(m-ny)] + \mathrm{erfi}[(1+j)(m+ny)]\}, \tag{21}$$

where the constants are $k = 1.05839(1-j) \times 10^{-4}$, $l = 4.34991 \times 10^8$, $m = 219.922$, and $n = 16102.6$. The magnitude and phase of Eq. (21) are illustrated in Figs. 3(a) and 3(b), respectively. Figure 3(b) shows only the central tenth of the phase. This is to facilitate direct comparison with the discrete approximations only in the unwrapped segment of the phase, as use of phase unwrapping software would introduce errors that are not of interest in this paper.

In Fig. 4, we compare Eq. (21) with the numerical approximation of the equation performed using a DLCT in MATLAB, using 2048 samples. Figure 4(a) shows the phase error over the same range as before. The phase er-

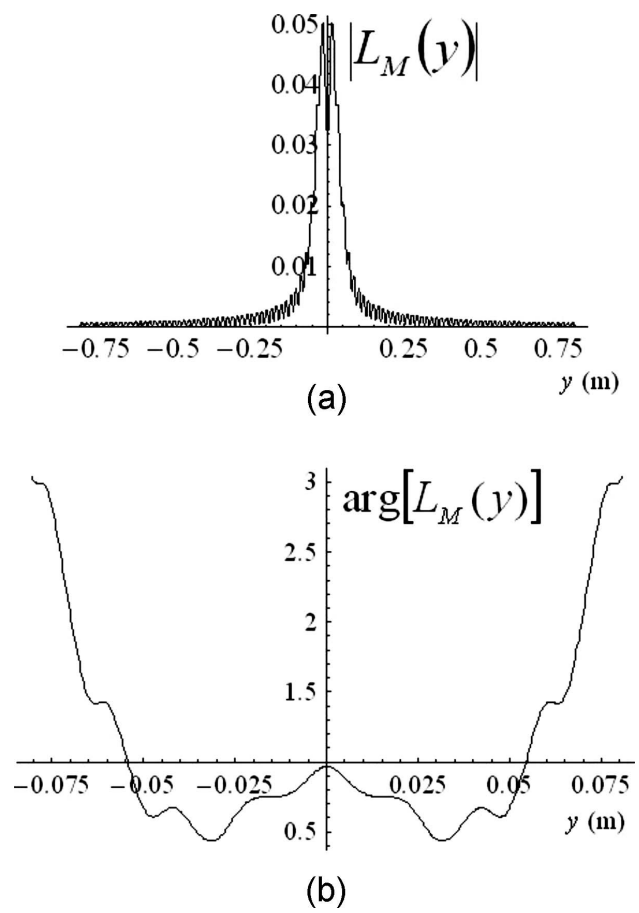

Fig. 3. (a) Amplitude and (b) wrapped phase of the LCT for parameters $\alpha = 636.619249$, $\beta = 636.620034$, and $\gamma = 636.619249$ of a 1D rectangular aperture of width $w = 9.96$ cm, determined analytically. $\lambda = 1$.

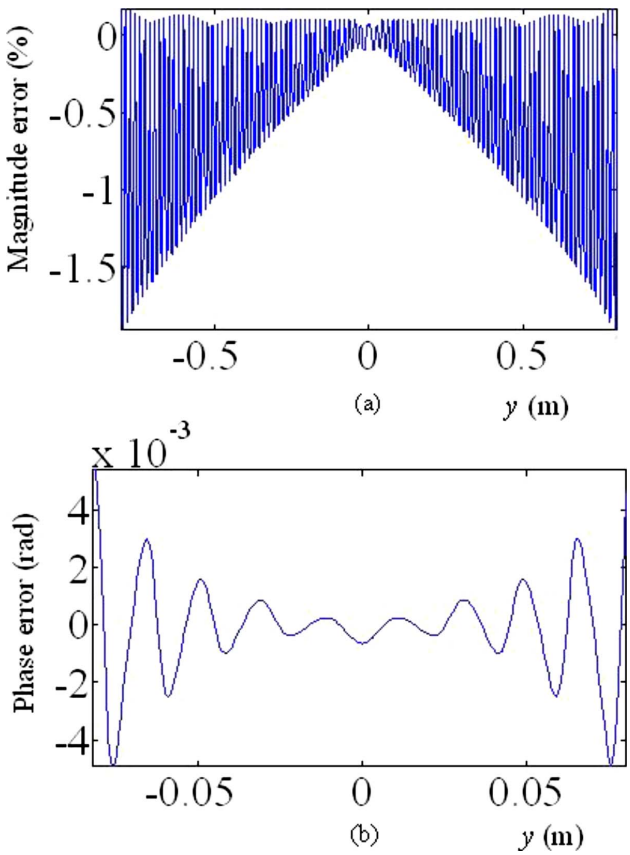

Fig. 4. (Color online) Comparison between the analytic result of Fig. 4 and the same calculation approximated using the discrete LCT. (a) Magnitude error, expressed as a percentage of the value at zero of the discrete output. (b) Phase error.

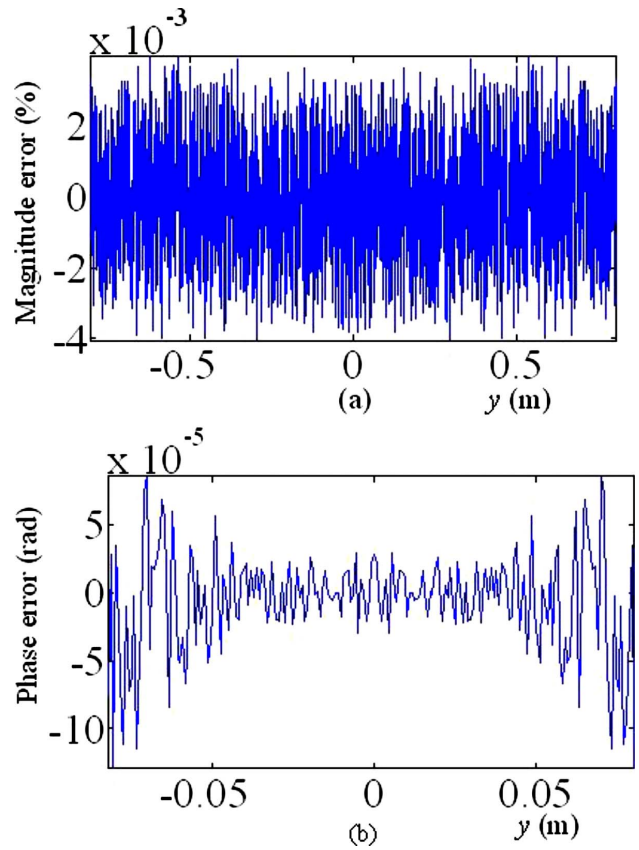

Fig. 5. (Color online) Comparison between the result of Fig. 4 performed using $O(N^2)$ and $O(N \log N)$ algorithms ($N=2048$). (a) Phase error. (b) Magnitude error, expressed as a percentage of the value at zero of the discrete output.

ror is of the order of 10^{-4} rad. We note that increasing the sampling rate used or zero padding the input reduced this error. Figure 4(b) shows the magnitude error as a percentage of the $y=0$ term of the DLCT result. In the central region, where most of the power is concentrated, the error is of the order of 0.1%. This grows toward the edges to about 1.75%. Displaying the discrete and analytic results on the same graph, we see that the discrete result is tending to zero faster than the analytic function. If we were to calculate the LCT over a wider window, we would expect to see the error saturate as the discrete function approached zero and eventually decrease as the analytic function also approached zero. Adjusting the input slightly presents cases where the period of the sidelobes differs between the discrete and analytic cases, resulting in a less structured error. However, there is consistently some error toward the edges of the result. If we compare the analytic Fourier transform of the input rect function with the output of MATLAB's FFT function, we see similar results, and we conclude that the discrepancies are due to aliasing. We conclude that if the DFT satisfactorally approximates the FT, then the DLCT satisfactorily approximates the LCT.

In order to ascertain whether the algorithm produces unacceptable errors, Fig. 5 compares the corresponding solutions resulting from the $O(N^2)$ DLCT algorithm with the radix-2 FLCT implementation derived in Section 4. Figure 4(a) shows the phase error. We note that the peak errors are of the order of 10^{-3} rad. Figure 4(b) shows the magnitude error as a percentage of the dc value of the output. We note that the peak errors are of the order of 0.1%. We conclude that the fast algorithm evaluates the DLCT with acceptably small deviations. It is unclear to what extent these errors are intrinsic to the algorithm and to what extent they depend on the implementation, but that discussion is beyond the scope of this paper.

6. CONCLUSION

We have reviewed in detail the literature pertaining to numerical approximation of the linear canonical transform (LCT) for simulation of quadratic phase systems. We have derived an $O(N \log N)$ algorithm of arbitrary radix for evaluating the discrete LCT (DLCT). We have presented the results of a radix-2 implementation of this fast LCT (FLCT) algorithm, confirming that it adequately evaluates the DLCT and also adequately approximates the continuous LCT.

Based on the current state of the literature, we believe that the theory of numerically approximating the LCT has reached a point comparable to that of the FT in 1965, when Cooley and Tukey published their celebrated paper [50]. The DLCT is now well defined, a reasonably mature theory exists for sampling and for the discrete form of the transformation, and early fast algorithms exist. To date, however, no prime-length algorithm is known. The field is

now well placed for computer scientists to tune the algorithms and develop specific implementations comparable to the extremes modern FFT implementations have reached, e.g., the eccentrically named "Fastest FFT in the West" (FFTW) [45]. The FLCT should be of immediate interest in optical-system-design software, where ray tracing dominates because of its flexibility and relative ease of computation. The improved speed of, and reduced computational requirements for, such scalar diffraction algorithms will further lower the barrier to more frequent use of such models. Furthermore, we note also that the existing theory may be modified to handle lossy systems (nonsymplectic ABCD matrices). As 40 years of research in digital signal processing and computer science has not exhausted the potential for further refinement of FFT algorithms, e.g., [51–53], it appears to be simply a matter of whether the FLCT can now attract the attention the FFT continues to receive.

As noted in the text, in relation to optical systems, one avenue of research that appears to have been neglected to date is the simulation of systems that are nonorthogonal, i.e., that are not separable in the two spatial dimensions and so cannot be modeled using the known 1D algorithms. It would be useful either to find a means of adapting the 1D algorithms to perform such operations or to derive a specifically 2D algorithm, starting with a 2D LCT that does not assume orthogonality. Another area of interest is the search for a fast algorithm for calculating the LCT using polar coordinates. The special case of this for the FT remains an area of active research [54].

ACKNOWLEDGMENTS

The authors acknowledge the support of Enterprise Ireland, Science Foundation Ireland, and the Irish Research Council for Science, Engineering and Technology. J. J. Healy acknowledges the support of FÁS, the Irish National Training and Employment Authority, through the FÁS Science Challenge.

REFERENCES

1. T. Kreis, *Handbook of Holographic Interferometry, Optical and Digital Methods* (Wiley-VCH, 2005).
2. J. W. Goodman, *Introduction to Fourier Optics*, 3rd ed. (Roberts, 2005).
3. S. Abe and J. T. Sheridan, "Optical operations on wave functions as the Abelian subgroups of the special affine Fourier transformation," Opt. Lett. **19**, 1801–1803 (1994).
4. H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing* (Wiley, 2001).
5. M. J. Bastiaans, "Wigner distribution function and its application to first-order optics," J. Opt. Soc. Am. **69**, 1710–1716 (1979).
6. H. O. Bartelt, K.-H. Brenner, and A. W. Lohmann, "The Wigner distribution function and its optical production," Opt. Commun. **32**, 32–38 (1980).
7. K. B. Wolf, *Integral Transforms in Science and Engineering* (Plenum, 1979).
8. S. A. Collins, "Lens-system diffraction integral written in terms of matrix optics," J. Opt. Soc. Am. **60**, 1168–1177 (1970).
9. S. Abe and J. T. Sheridan, "Generalization of the fractional Fourier transformation to an arbitrary linear lossless transformation an operator approach," J. Phys. A **27**, 4179–4187 (1994).
10. B. M. Hennelly, D. P. Kelly, R. F. Patten, J. E. Ward, U. Gopinathan, F. T. O'Neill, and J. T. Sheridan, "Metrology and the linear canonical transform," J. Mod. Opt. **53**, 2167–2186 (2006).
11. D. P. Kelly, J. E. Ward, B. M. Hennelly, U. Gopinathan, F. T. O'Neill, and J. T. Sheridan, "Paraxial speckle-based metrology systems with an aperture," J. Opt. Soc. Am. A **23**, 2861–2870 (2006).
12. R. F. Patten, B. M. Hennelly, D. P. Kelly, F. T. O'Neill, Y. Liu, and J. T. Sheridan, "Speckle photography: mixed domain fractional Fourier motion detection," Opt. Lett. **31**, 32–34 (2006).
13. B. M. Hennelly and J. T. Sheridan, "Image encryption and the fractional Fourier transform," Optik (Stuttgart) **114**, 251–265 (2003).
14. A. Nelleri, J. Joseph, and K. Singh, "Digital Fresnel field encryption for three-dimensional information security," Opt. Eng. (Bellingham) **46**, 045801 (8 pages) (2007).
15. J. E. Ward, D. P. Kelly, and J. T. Sheridan, "Three-dimensional speckle size in generalized optical systems with limiting apertures," J. Opt. Soc. Am. A **26**, 1858–1867 (2009).
16. M. J. Bastiaans and K. B. Wolf, "Phase reconstruction from intensity measurements in one-parameter canonical-transform systems," in *Proceedings of Seventh International Symposium on Signal Processing and Its Applications*, Vol. 1 (IEEE, 2003), pp. 589–592.
17. U. Gopinathan, G. Situ, T. J. Naughton, and J. T. Sheridan, "Noninterferometric phase retrieval using a fractional Fourier system," J. Opt. Soc. Am. A **25**, 108–115 (2008).
18. S.-C. Pei and J.-J. Ding, "Closed-form discrete fractional and affine Fourier transforms," IEEE Trans. Signal Process. **48**, 1338–1353 (2000).
19. B. M. Hennelly and J. T. Sheridan, "Fast numerical algorithm for the linear canonical transform," J. Opt. Soc. Am. A **22**, 928–937 (2005).
20. J. J. Healy and J. T. Sheridan, "Sampling and discretization of the linear canonical transform," Signal Process. **89**, 641–648 (2009).
21. J. J. Healy, B. M. Hennelly, and J. T. Sheridan, "An additional sampling criterion for the linear canonical transform," Opt. Lett. **33**, 2599–2601 (2008).
22. A. Stern, "Why is the linear canonical transform so little known?" in *Proceedings of 5th International Workshop on Information Optics*, G. Cristóbal, B. Javidi, and S. Vallmitjana, eds (Springer, 2006), pp. 225–234.
23. F. Oktem and H. M. Ozaktas, "Exact relation between continuous and discrete linear canonical transforms," IEEE Signal Process. Lett. **16**, 727–730 (2009).
24. A. Koç, H. M. Ozaktas, C. Candan, and M. A. Kutay, "Digital computation of linear canonical transforms," IEEE Trans. Signal Process. **56**, 2383–2394 (2008).
25. B. M. Hennelly and J. T. Sheridan, "Generalizing, optimizing, and inventing numerical algorithms for the fractional Fourier, Fresnel, and linear canonical transforms," J. Opt. Soc. Am. A **22**, 917–927 (2005).
26. X. Liu and K.-H. Brenner, "Minimal optical decomposition of ray transfer matrices," Appl. Opt. **47**, E88–E98 (2008).
27. T. Alieva and M. J. Bastiaans, "Properties of the linear canonical integral transformation," J. Opt. Soc. Am. A **24**, 3658–3665 (2007).
28. G. Kloos, *Matrix Methods for Optical Layout* (SPIE Press, 2007).
29. L. Onural, "Some mathematical properties of the uniformly sampled quadratic phase function and associated issues in digital Fresnel diffraction simulations," Opt. Eng. (Bellingham) **43**, 2557–2563 (2004).
30. F. Gori, "Fresnel transform and sampling theorem," Opt. Commun. **39**, 293–297 (1981).
31. J. J. Healy and J. T. Sheridan, "Cases where the linear canonical transform of a signal has compact support or is band-limited," Opt. Lett. **33**, 228–230 (2008).
32. X.-G. Xia, "On bandlimited signals with fractional Fourier

transform," IEEE Trans. Signal Process. **3**, 72–74 (March 1996).
33. J.-J. Ding, "Research of fractional Fourier transform and linear canonical transform," Ph.D. dissertation (National Taiwan University, 2001).
34. A. Stern, "Sampling of linear canonical transformed signals," Signal Process. **86**, 1421–1425 (2006).
35. B. Deng, R. Tao, and Y. Wang, "Convolution theorems for the linear canonical transform and their applications," Sci. China Ser. F, Inf. Sci. **49**, 592–603 (2006).
36. B.-Z. Li, R. Tao, and Y. Wang, "New sampling formulae related to linear canonical transform," Signal Process. **87**, 983–990 (2007).
37. J. Zhao, R. Tao, and Y. Wang, "Sampling rate conversion for linear canonical transform," Signal Process. **88**, 2825–2832 (2008).
38. B. Barshan, M. Alper Kutay, and H. M. Ozaktas, "Optimal filtering with linear canonical transformations," Opt. Commun. **135**, 32–36 (1997).
39. A. W. Lohmann, R. G. Dorsch, D. Mendlovic, Z. Zalevsky, and C. Ferreira, "Space–bandwidth product of optical signals and systems," J. Opt. Soc. Am. A **13**, 470–473 (1996).
40. W. T. Rhodes, "Light tubes, Wigner diagrams and optical signal propagation simulation," in *Optical Information Processing: A Tribute to Adolf Lohmann*, H. J. Caulfield, ed. (SPIE Press, 2002) pp. 343–356 (2002).
41. J.-C. Kuo, C.-H. Wen, and A.-Y. Wu, "Implementation of a programmable 64∼2048-point FFT/IFFT processor for OFDM-based communication systems," in *Proceedings of the 2003 International Symposium on Circuits and Systems*, Vol. 2 (IEEE, 2003), pp. 121–124.
42. H. M. Ozaktas, A. Koç. I. Sari, and M. Alper Kutay, "Efficient computation of quadratic-phase integrals in optics," Appl. Opt. **31**, 35–37 (2006).
43. H. M. Ozaktas, O. Arikan, M. A. Kutay, and G. Bozdagt, "Digital computation of the fractional Fourier transform," IEEE Trans. Signal Process. **44**, 2141–2150 (1996).
44. W. M. Gentleman and G. Sande, "Fast Fourier transforms—for fun and profit," in *Proceedings of AFIPS Fall Joint Computer Conference*, Vol. 29 (ACM, 1966), pp. 563–578.
45. M. Frigo and S. G. Johnson, "The design and implementation of FFTW3," Proc. IEEE **93**, 216–231 (2005).
46. C. M. Rader, "Discrete Fourier transforms when the number of data samples is prime," Proc. IEEE **56**, 1107–1108 (1968).
47. B. Kernighan, and D. Ritchie, *The C Programming Language* (Prentice-Hall, 1978).
48. MATLAB, *Users Guide* (The MathWorks, Inc., 1998).
49. S. Wolfram, *The Mathematica Book*, 3rd ed. (Wolfram Media, 2003).
50. J. W. Cooley and J. W. Tukey, "An algorithm for the machine computation of complex Fourier series," Math. Comput. **19**, 297–301 (1965).
51. S. G. Johnson and M. Frigo, "A modified split-radix FFT with fewer arithmetic operations," IEEE Trans. Signal Process. **55**, 111–119 (2007).
52. T. J. Lundy and J. Van Buskirk, "A new matrix approach to real FFTs and convolutions of length 2^k," Computing **80**, 23–45 (2007).
53. A. Cortes, I. Velez, and J. F. Sevillano, "Radix r^k FFTs: matricial representation and SDC/SDF pipeline implementation," IEEE Trans. Signal Process. **57**, 2824–2839 (2009).
54. A. Averbuch, R. R. Coifman, D. L. Donoho, M. Elad, and M. Israeli, "Fast and accurate polar Fourier transform," Appl. Comput. Harmon. Anal. **21**, 145–167 (2006).