Frequency-based Environment Matting by Compressive Sensing

Yiming Qian
University of Alberta
yqian3@ualberta.ca

Minglun Gong
Memorial University of Newfoundland
gong@cs.mun.ca

Yee-Hong Yang
University of Alberta
yang@cs.ualberta.ca

Abstract

Extracting environment mattes using existing approaches often requires either thousands of captured images or a long processing time, or both. In this paper, we propose a novel approach to capturing and extracting the matte of a real scene effectively and efficiently. Grown out of the traditional frequency-based signal analysis, our approach can accurately locate contributing sources. By exploiting the recently developed compressive sensing theory, we simplify the data acquisition process of frequency-based environment matting. Incorporating phase information in a frequency signal into data acquisition further accelerates the matte extraction procedure. Compared with the state-of-the-art method, our approach achieves superior performance on both synthetic and real data, while consuming only a fraction of the processing time.

1. Introduction

Environment matting is a technique introduced to render photo-realistic images of objects that are either refractive or reflective, or both. It focuses on modeling the object's light transport characteristics, *i.e*. the *environment matte*, and allows the object to be seamlessly composited into a new background. Typically, to obtain an environment matte, the object needs to be photographed in front of a series of predesigned backdrops. How the object refracts and reflects light can then be inferred from the recorded images.

The concept of environment matting is first introduced in [20]. Since then, several methods [5, 7, 13, 17, 19] have been proposed to either simplify the data acquisition process or to improve the accuracy of the environment matte. The main task of environment matting is to decompose a *many-to-one* mapping, by which many background pixels are combined into one foreground pixel. Most existing methods decompose the mapping in the spatial domain, where a foreground pixel can be composited in an infinite number of ways. To handle the ambiguities, these methods use additional constraints and time-consuming non-linear optimization to estimate the mapping, whose physical correctness cannot be verified. A frequency-based approach is later proposed and is capable of finding the accurate contributing sources efficiently [19]. However, it requires a large number of captured images. On the other hand, extracting matte data from a single photo is possible if the object is perfectly specular transparent [5]. A recent spatial-domain method [7] also has low data acquisition requirement at the expense of high computational cost. Both of these two methods cannot provide high-accuracy environment mattes. Hence, there is a need for an effective algorithm that can accurately and quickly extract the matte data from a small number of images.

Motivated by the above observations, in this paper, we present a novel environment matting approach with the following objectives. First, our approach can accurately find the contributing sources at the pixel level. Second, our approach leverages on the recently developed theory of *Compressive Sensing (CS)* to reduce the complexity of data acquisition. Finally, we incorporate additional phase information into the frequency-based model, which further reduces the number of images required and significantly accelerates the process of environment matte extraction.

2. Related Work

Environment Matting. Zongker *et al*. [20] first formulate the problem and decompose the many-to-one mapping by assuming a foreground pixel is only contributed by a single rectangle area of the background. Chuang *et al*. [5] later propose two extensions: i) by sweeping different oriented Gaussian strips across the background to accommodate for sources that are not axis-aligned rectangles; ii) by extracting the environment mattes of colorless and pure specular objects using only one image. Wexler *et al*. [17] present a probabilistic model based method, which does not rely on predefined backdrops but requires enough sample images. It only works for thin transparent objects that do not introduce large optical distortion to the background. Inspired by image-based relighting, Peers and Dutr [13] use a set of wavelet basis images to obtain a visually pleasing result using a large number of sample images.

Inspired by the fact that a signal has a unique decomposi-

tion in the frequency domain, Zhu and Yang [19] propose a frequency-based approach which can find the accurate contributing sources, allowing the decomposition ambiguity to be alleviated. Our proposed approach is built upon their frequency-based model, but uses CS to dramatically reduce the number of images required.

Note that using CS-based data acquisition for environment matting has been recently proposed by Duan *et al.* [7, 8]. Our work differs from theirs in three main aspects: 1) Rather than solving the problem in the spatial domain, we work in the frequency domain, which helps to accurately locate the contributing sources; 2) The sparsity assumptions in CS are different. Their method assumes that the light transport vector is sparse and directly reconstructs it in the spatial domain, whereas we assume a foreground pixel is contributed by a sparse number of frequencies, not locations; 3) To reduce the computational cost, the hierarchical recovery scheme they proposed limits the contributions to a foreground pixel by only square blocks in the background. Hence, their composited results appear blurry and blocky. In contrast, our results are sharper and clearer because our approach can locate the contributing sources at the pixel level efficiently.

Compressive Sensing. Compressive sensing [3, 6] is an emerging field that provides a framework for reconstructing a sparse signal with far fewer measurements than the dimension of the signal. Instead of capturing the original N-dimensional signal x directly, to recover x, CS seeks to use $M < N$ linear measurements $y = Ax$, where A is an $M \times N$ measurement matrix, and x is an s-sparse signal, *i.e.* x contains at most $s \ll N$ nonzero elements. In CS, if the measurement matrix A satisfies the restricted isometry property (RIP) [2], then x can be stably recovered by solving the nonlinear optimization problem: $\min \|x\|_1$, s.t. $y = Ax$ with only $M = \mathcal{O}(s \log(N/s))$ measurements.

CS has facilitated the solving of many computer vision and graphics problems, either by helping in reformulating the problem using the sparsity constraint, *e.g.* face recognition [18], background subtraction [4], or by reducing the complexity of data acquisition, *e.g.* light transport acquisition [14], dual photography [16]. As mentioned before, CS is also incorporated into environment matting in the spatial domain [7]. Huang *et al.* [9] propose a CS-based solution for recovering data with both sparsity and dynamic group clustering priors. Note that the group clustering prior is also applicable to environment matting and has been utilized in [7] since the background pixels contributed to an object pixel appear in groups.

3. Prerequisites

Problem Formulation. An environment matte describes how light is transferred from the environment through a transparent or reflective object to the camera. Following [7, 13, 19], the problem is usually modeled as

$$C = F + \rho \mathbf{W}\mathbf{B}, \tag{1}$$

where C is the intensity of a pixel in the composited image, and F the ambient illumination. $\mathbf{B}$ is an $n^2 \times 1$ vector representing the background image, and $\mathbf{W}$ the $1 \times n^2$ light transport vector describing the amount of contribution of light emitted from each background pixel to an object pixel, with the constraints $\|\mathbf{W}\|_1 = 1, \mathbf{W}_i \geq 0$. ρ is the light attenuation index which defines how light is attenuated by the object. In this way, each object pixel C is a combination of the ambient illumination F and the weighted contribution of the light emitted from the backdrop $\mathbf{B}$. Hence, the problem becomes: *given a number of captured images of an object against some known backdrops, how to extract the environment mattes: F, ρ and $\mathbf{W}$?*

Previous methods have shown that obtaining F and ρ under controlled environment is relatively easy [19, 20]. In particular, F can be obtained by displaying a pure black background because $C = F$ when there is no background contribution. ρ can be obtained by projecting a solid color background, where all entries in $\mathbf{B}$ have the same value b. Consider $\|\mathbf{W}\|_1 = 1$, Eq.(1) becomes $C = F + \rho b$, allowing ρ to be calculated after F is determined.

Thus the main task of environment matting is to recover the light transport vector $\mathbf{W}$ for each pixel. It is worth noting that we are only interested in the foreground object pixels, which are specified using a binary mask. The mask is obtained by capturing the scene with and without the object in front of 20 coarse-to-fine backdrops [20]. A pixel is considered an object pixel if the corresponding pixel colors of the object image and the reference image differ by more than a threshold in any of the 20 pairs. Two subsequent morphological operations, an opening followed by a closing operation with a 5×5 box structural element, are used to further refine the mask.

Frequency Analysis Model. In an effort to alleviate the ambiguity problem, Zhu and Yang [19] propose to estimate the matte in the frequency domain. The key idea is to utilize the following desirable properties of the *Discrete Fourier Transform (DFT)*:

1. Suppose a signal s_3 is a weighted combination of two other signals s_1 and s_2, *i.e.* $s_3 = w_1 s_1 + w_2 s_2$. Denote the frequency of s_1 and s_2 as f_1 and f_2, respectively, then s_3 is a signal with both f_1 and f_2;

2. Denote the complex vector S_3 as the DFT of s_3, we have $\mathrm{mag}(S_3(f_1)) > 0$, $\mathrm{mag}(S_3(f_2)) > 0$ and $\frac{\mathrm{mag}(S_3(f_1))}{\mathrm{mag}(S_3(f_2))} = \frac{w_1}{w_2}$, where $\mathrm{mag}(\cdot)$ denotes the complex magnitude of a complex number;

3. DFT is robust to noise.

By letting different pixels emit different frequency signals in the backdrop, we can apply the DFT to the observed signal of each object pixel and find the peaks of the frequency magnitude, which correspond to the contributing sources in the backdrop. The weights of these sources, *i.e*. the vector $\mathbf{W}$, can then be computed using the aforementioned property 2. However, for a backdrop with n^2 pixels ($n \approx 10^3$ for a conventional monitor), assigning each pixel a unique frequency requires at least $2 \times n^2$ images to be captured so that the frequency information can be recovered based on the Nyquist-Shannon Sampling Theorem. Capturing so many images is impractical and time-consuming.

To reduce the number of captured images, Zhu and Yang [19] split the data acquisition into two stages. Row-based patterns are first captured, where pixels in a row have the same frequency, then column-based patterns are captured, where pixels in a column share the same frequency. The final contributing sources can be jointly determined by row-based and column-based searching. While the number of images needed is reduced from $2 \times n^2$ to $4 \times n$, thousands of images are still needed to extract the matte at pixel level using a typical monitor. In addition, this encoding scheme assumes that the contributing sources can be depicted using the element-wise product of a row vector and a column vector, which may not hold for all objects; see Figure 6 and Section 6 for detailed discussions on this limitation.

4. Proposed Approach

In this section, we first present the sparsity of contributing sources under the frequency-based formulation. Then a CS-based reconstruction method is introduced to simplify the data acquisition process. Finally, we present a novel background design with phase incorporated to reduce the computational cost of L_1 minimization in CS.

4.1. Sparsity under Frequency-based Formulation

Unlike the previous work [7] that assumes the sparsity of the light transport vector $\mathbf{W}$ in Eq.(1), here we show the sparsity of contributing sources in our frequency-based pattern configuration. In particular, when row-based or column-based backdrops are displayed, an object pixel is only contributed by a few rows or columns. Hence, the corresponding DFT contains a small number of frequencies.

To quantitatively justify the sparsity of contributing frequencies in the recorded signal, we capture several objects under row-based and column-based frequency patterns. For row-based patterns, the intensity of each row in the temporal sequence is designed as

$$B(f,t) = \xi\Big(\cos\big(2\pi f \frac{t}{N}\big) + 1\Big), \tag{2}$$

where $1 \leq f \leq n$ is the row index of the background image, which also represents the frequency value at the fth row.

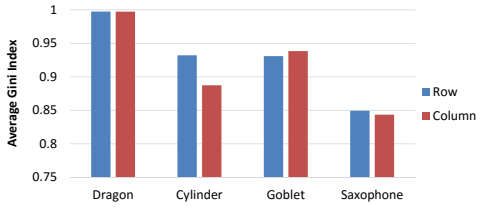


Figure 1. The Gini indices of the four objects used in this paper. The resolution of the row-based and column-based pattern is set to 512×512 here for practical data acquisition.

ξ is set to 127.5 such that the range of pixel values is in $[0, 255]$. N represents the inverse of the sampling period, and we have $N \geq 2f_{max}$ according to the Nyquist-Shannon Sampling Theorem. In practice, we set $N = 2f_{max} + o$ and $o \in [10, 15]$ is an offset term. t is the time index (frame id) within the set $\{0, 1, \cdots, N-1\}$. The column-based pattern follows the same fashion.

After recording the object images under two kinds of backdrops, we apply the DFT to the received signals at each foreground pixel. For a single pixel under the row-based pattern, if the complex number at frequency f of the DFT is non-zero (the corresponding magnitude is non-zero), it means that the received signal contains that frequency. Thus we conclude that the fth row in the background contributes to the pixel. Hence, the complex magnitudes of the DFT can be used to measure the sparsity of frequencies of each object pixel. We compute the Gini indices [10] (a widely-used sparsity metric in signal processing, with a higher value implying a sparser signal) of the magnitude vectors for all object pixels under both row-based and column-based patterns, and average them as the sparsity of rows and columns, respectively. As shown in Figure 1, the test objects have consistently high sparsity for both rows and columns. In the following subsections, we use the row-based pattern to illustrate our approach, since the column-based pattern is analyzed in the same fashion.

4.2. Reconstruction via Compressive Sensing

The existing frequency-based approach requires at least $4n$ (4096 when $n = 1024$) captured images for both row and column-based patterns. In contrast, the proposed CS-based approach utilizes the sparsity in frequencies to reduce the number of images required. To derive our CS-based method, we first consider the conventional DFT method for reconstructing frequency information. Given the recorded signal $\mathbf{C}$ and the computed ambient illumination F of a foreground pixel, we have $\mathbf{C} - F = \mathbf{DX}$, where $\mathbf{X}$ is an N-dimensional complex vector representing the frequency information of an object pixel and note that during searching non-zero frequencies, we are only interested in the sub-vector $\mathbf{X}(1 : f_{max})$. $\mathbf{D}$ is the inverse of the $N \times N$ discrete Fourier transform matrix. Since each object pixel is contributed by only a few rows, *i.e*. frequencies, the CS theory

can be used to reconstruct the sparse frequency information $\mathbf{X}$ by taking only $M < N$ measurements.

In practice, by randomly generating an M-dimensional permutation Ω of the set $\{0, 1, \cdots, N-1\}$ and displaying M backdrops pre-computed using Eq.(2) with frame ids from Ω, we can solve the following L_1 minimization problem to reconstruct the frequency information $\mathbf{X}$:

$$\min \|\mathbf{X}\|_1, \text{ s.t. } \mathbf{C} - F = \mathbf{D}(\Omega, :)\mathbf{X}, \tag{3}$$

where $\mathbf{C}$ is an M-dimensional vector representing the recorded temporal signal of a foreground pixel, and $\mathbf{D}(\Omega, :)$ is the measurement matrix extracted from $\mathbf{D}$ by including only the rows with indices in Ω.

Besides sparsity, previous works [5, 7, 17] have shown the background regions that contribute to a foreground pixel can be clustered into several main groups. Since the signal frequencies correlate with pixel locations, such a group prior also exists in our frequency-based formulation. That is, most non-zero elements in $\mathbf{X}$ are neighbors and can be clustered into several local groups, which can help to improve the accuracy of L_1 minimization in Eq.(3). In practice, we apply the DGS tool [9] that can automatically handle the group clustering prior during optimization.

After $\mathbf{X}$ is obtained, a simple thresholding operation is performed to locate the contributing rows, *i.e.* the frequencies with non-zero magnitude. The threshold is set as $\max(\text{mag}(\mathbf{X}))/2$ in our implementation. Together with the contributing columns located in the same manner, the locations of contributing sources are thereby determined. The weight of the source at row r and column c is calculated as

$$\mathbf{W}(\text{ind}(r, c)) = \bar{\mathbf{W}}_{row}(r)\bar{\mathbf{W}}_{col}(c), \tag{4}$$

where $\text{ind}(\cdot, \cdot)$ returns the 1D index of the pixel located at the rth row and cth column in the background. $\mathbf{W}_{row}$ and $\mathbf{W}_{col}$ are computed from the frequency information $\mathbf{X}$ of row-based and column-based acquisitions, respectively, and are normalized before plugging into Eq.(4). Figure 3(c) is an example using the proposed CS-based frequency reconstruction, where $M = 160$ and $N = 2085$ are used for both row-based and column-based acquisitions.

4.3. Augment with Phase Information

The CS-based frequency search lowers the data acquisition requirement, but at the cost of a more expensive reconstruction process. Since the details of composition results depend on the resolution of the background pattern n, n needs to be large enough. When $n = 1024$, then the maximal frequency $f_{max} = 1024$ in the row-based and column-based patterns. Hence, the unknown vector $\mathbf{X}$ has a dimension of $N \geq 2f_{max} = 2048$. Solving such a large constrained minimization problem for all foreground pixels is time-consuming, *e.g.* extracting the environment matte of the object in Figure 3(c) takes over 26 minutes.

4.3.1 Background Pattern Design

To accelerate the process of solving L_1 minimization, we develop a new method that incorporates additional phase information to reduce the complexity of minimizing the L_1 norm. The core idea is to use both frequency and phase to identify the contributing sources. That is, for row-based patterns, we split the image into k horizontal regions. While different rows within the same region all have different frequencies, the corresponding rows in different regions have the same frequency but different phase values. This is achieved by assigning pixels in the fth row of the pth region the intensity of

$$B(f, t, \varphi_p) = \xi\Big(\cos\big(2\pi f\frac{t}{N} + \varphi_p\big) + 1\Big), \tag{5}$$

where $1 \leq f \leq \frac{n}{k}$ is the row index in the pth region, which also represents the corresponding frequency value. φ_p is a pre-designed phase value for the pth ($1 \leq p \leq k$) region. How to properly assign φ_p is discussed in Section 4.3.3.

Adding to the background with k phases reduces the maximum frequency requirement from n to $\frac{n}{k}$, which subsequently reduces the dimension of $\mathbf{X}$ by k times and the computational cost of the L_1 minimization in Eq.(3). On the other hand, to determine the contributing sources, we need both frequency search and phase search. In practice, given the recorded temporal signal at a foreground pixel, we first determine the frequencies of the contributing sources by optimizing Eq.(3). Then, for a contributing frequency f, we compute its phase value to locate the region from which the frequency originates. Combining the phase and the frequency information gives us the row index in the background image.

4.3.2 Phase Acquisition and Inference

According to the theory of the DFT, given a set of phase candidates, the complex number $\mathbf{X}(f)$ is a weighted combination of different phase data:

$$\mathbf{X}(f) = \sum_{p=1}^{k} \alpha_p \left(\cos\varphi_p + j\sin\varphi_p\right) = R + jI, \tag{6}$$

where R and I are, respectively, the known real and imaginary part of $\mathbf{X}(f)$. If the frequency f comes from the pth region, we should have $\alpha_p > 0$ and vice versa. Therefore, if we know the coefficients α's, the contributing sources can be easily located. Considering the real and imaginary parts of Eq.(6) separately, we have two equalities. Hence, when $k = 2$, the two coefficients, α_1 and α_2, can be directly solved. When $k > 2$, additional equalities are required to compute the k coefficients.

To address the problem, we capture more frequency-based patterns under different phase settings. It is noteworthy that, regardless of the setting of the phase candidates $\{\varphi_1, \cdots, \varphi_k\}$, the complex number $\mathbf{X}(f)$ is non-zero

as long as the fth row in some regions makes contribution to the object pixel. Furthermore, the coefficients α's are independent of the phase setting since they represent the amount of light from different regions. Therefore, to obtain k equalities for solving the k coefficients, we have to capture row-based patterns generated from Eq.(5) using $\frac{k}{2}$ different phase settings. It is worth noting that, due to the needs for additional phase setting, once $k > 2$, increasing k no longer reduces the number of background images needed. Nevertheless, the benefit of reducing the dimension of $\mathbf{X}$ remains.

Denote each phase setting as $\{\varphi_p^q : 1 \le p \le k\}$, where p is the region index and $1 \le q \le \frac{k}{2}$ the phase setting index. For each phase setting, by capturing the corresponding row-based patterns, we solve the optimization problem Eq.(3) to recover the frequencies. Then for each frequency, we have $\frac{k}{2}$ complex numbers: $\mathbf{X}^1(f) = R^1 + jI^1, \cdots, \mathbf{X}^{\frac{k}{2}}(f) = R^{\frac{k}{2}} + jI^{\frac{k}{2}}$, and they satisfy Eq.(6). Considering the real and imaginary parts separately, we have:

$$\begin{bmatrix} \cos\varphi_1^1 & \cos\varphi_2^1 & \cdots & \cos\varphi_k^1 \\ \vdots & \vdots & \ddots & \vdots \\ \cos\varphi_1^{\frac{k}{2}} & \cos\varphi_2^{\frac{k}{2}} & \cdots & \cos\varphi_k^{\frac{k}{2}} \\ \sin\varphi_1^1 & \sin\varphi_2^1 & \cdots & \sin\varphi_k^1 \\ \vdots & \vdots & \ddots & \vdots \\ \sin\varphi_1^{\frac{k}{2}} & \sin\varphi_2^{\frac{k}{2}} & \cdots & \sin\varphi_k^{\frac{k}{2}} \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} = \begin{bmatrix} R^1 \\ \vdots \\ R^{\frac{k}{2}} \\ I^1 \\ \vdots \\ I^{\frac{k}{2}} \end{bmatrix}. \quad (7)$$

For each frequency f, the corresponding coefficients $\{\alpha_1(f), \cdots, \alpha_k(f)\}$ can be obtained by solving the above linear system. We say that the fth row in the pth region (*i.e.* the row indexed at $r = f + \frac{n}{k}(p-1)$ in the background) makes contribution to the foreground pixel iff $\alpha_p(f) > 0$. The weight of the rth row is $\bar{\mathbf{W}}_{row}(r) = \frac{\alpha_p(f)}{\sum_{p,f}\alpha_p(f)}$.

In practice, because of measurement noise, the rth row is considered as a contributing row only if $\bar{\mathbf{W}}_{row}(r) > T$, where $T = \max(\bar{\mathbf{W}}_{row}(r))/2$ is used in all our implementation. By splitting the column-based pattern along the column direction and following the same phase settings, the column weights can be obtained in the same fashion, then the light transport vector $\mathbf{W}$ is computed using Eq.(4).

4.3.3 Construction of Phase Settings

Denote the linear system Eq.(7) as $\mathbf{Q}\alpha = \epsilon$, which could be indeterminate when $\mathbf{Q}$ is singular because of inappropriate phase settings. In addition, if similar degrees are used for neighboring phases, phase inference using Eq.(7) could locate undesired regions because of measurement noise.

To construct valid phase settings, three rules need to be followed: 1) The constructed $\mathbf{Q}$ is non-singular; 2) To make each region distinguishable by phase, there are no duplicated degrees in each phase setting; 3) The difference in phase values of adjacent regions should be large enough ($|\cos\varphi_p - \cos\varphi_{p+1}| > 0.5$ is used in our implementation). Note that the phase values are in the range $[0, 360)$. In our implementation, we randomly generate phase settings from $\{0, 20, \cdots, 340\}$ until the three rules are satisfied.

Considering $\alpha \ge 0$, solving $\mathbf{Q}\alpha = \epsilon$ is a classical *Non-negative Least Squares (NLS)* problem [12]. Here we propose to apply L_1 regularization to solve the linear system, which is more robust to noise than NLS. In particular, we compute the coefficients α's by solving

$$\min \|\alpha\|_1, \text{ s.t. } \mathbf{Q}\alpha = \epsilon, \alpha \ge 0. \quad (8)$$

5. Experiments

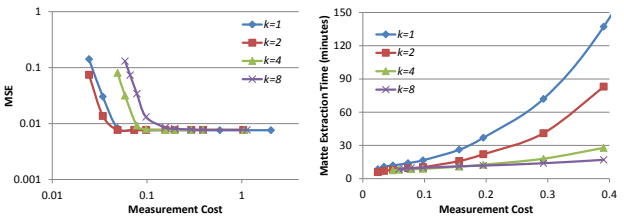
The proposed approach is tested using both synthetic and real transparent objects. The resolution of the background pattern is set to $n = 1024$ in all tests. Note that our CS-based data acquisition uses non-adaptive background patterns, which are generated and stored in advance. To prevent the interference caused by the bleeding effect of the monitor or other unknown light sources [19], the frequency range used is shifted up by 10Hz, *i.e.* $11 \le f \le 10 + \frac{n}{k}$. L_1 minimization is solved using the DGS tool [9] with group priors for Eq.(3) and without group priors for Eq.(8). Since the environment matte extraction process at each foreground pixel is independent, our parallel algorithm is implemented in MATLAB R2014b and accelerated on an 8-core PC with 3.4GHz Intel Core i7 CPU and 24GB RAM.

5.1. Synthetic Object

We start with quantitatively evaluate our approach using a complex synthetic model, the Stanford dragon, with our frequency-based backdrops texture mapped to the background. The data acquisition process is simulated using POV-Ray [1] and hence is free from measurement noise or lens imperfection. An image of the dragon in front of a checkerboard background is also rendered, which serves as the ground truth for measuring the *mean square errors* (MSE) of composited results.

Effectiveness of CS-based Phase-augmented Acquisition. We first evaluate the efficiency of the proposed data acquisition approach, which is usually quantified using the *measurement cost*, *i.e.* the ratio between the number of measurements and the number of unknowns. Here for each foreground pixel, we need to compute $\bar{\mathbf{W}}_{row}$ and $\bar{\mathbf{W}}_{col}$, which have a total of $2n$ unknowns. Denoting the total number of images used for both row and column based patterns as m, the measurement cost is defined as $\sigma = m/2n$.

As discussed in Section 3, the conventional frequency-based environment matting approach [19] requires $4n$ images. Hence, it has a measurement cost of 2. In our approach, with the additional phase search step, we only need to capture $2n$ images to reconstruct the frequency information using the DFT (*e.g.* in row-based acquisition, we have



(a) Accuracy (b) Matte Extraction Time

Figure 2. The impact of phase region number k and measurement costs on both accuracy of the composited result (a) and matte extraction time (b). Note that log scale is used for both axes in (a).

$k/2$ phase settings and for each phase setting we need to capture $2n/k$ images, thus n images are required). The measurement cost is therefore reduced to 1. Using CS-based acquisition can further lower the measurement cost dramatically without noticeably affecting the accuracy of the composited results. Figure 2(a) shows that the MSE remains to be low (< 0.01) when σ is set to about 0.1.

Figure 2(a) also shows the impact of the phase number k. As expected, without using phase ($k = 1$), a higher measurement cost is needed to achieve the same MSE than setting $k = 2$. Further increasing k results in more images needed to achieve a similar MSE. This is because using more phase regions corresponds to a fewer number of rows (columns) within the region. Since a foreground pixel is contributed by a fixed number of rows (columns), which often come from the same phase region, the frequency information becomes less sparse, which requires more samples for reconstruction according to the theory of CS.

Figure 2(b) further illustrates the impact of phases on the time of environment matte extraction. By splitting the pattern into more regions, the number of unknown frequencies decreases in Eq.(3), which accelerates the process of L_1 minimization, and thus speedups the whole process.

In summary, augmenting phase information can reduce the number of required images and accelerate the process of L_1 minimization. The phase region number k offers a tradeoff between the process of data acquisition and matte extraction. If the goal is to minimize the measurement cost while maintaining the accuracy of the composited results, $k = 2$ is the optimal setting. If the computational resource is limited, then a large k value should be used, which helps to reduce the cost of CS-based reconstruction.

Comparisons with the CS in Spatial Domain. Finally, we compare our composited results with the latest CS-based environment matting method [7], which solves the problem in the spatial domain. The method of [7] is implemented by us based on their paper. To accelerate matte extraction, their method splits background pattern into square blocks and assumes a foreground pixel is contributed by these blocks. Thus their results appear blurry and blocky. As shown in Figure 3, our approach achieves superior performance in terms of both MSE and matte extraction time.

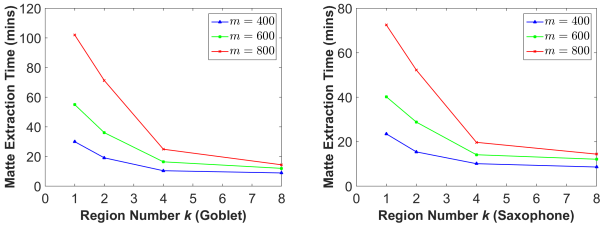


Figure 5. Effectiveness of using phase for accelerating matte extraction for two real objects: Goblet and Saxophone. Both results show that setting $k = 4$ can lower the computation cost by several manifolds, especially when a large number of captured images (m) is used. The benefit of further increasing k to 8 is limited.

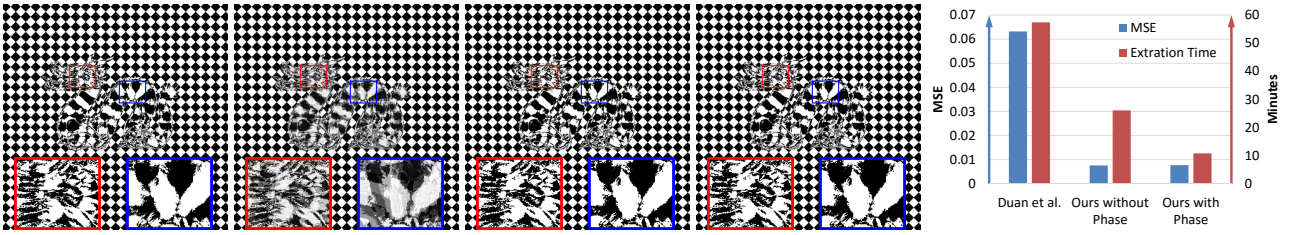
5.2. Real Transparent Objects

Five objects, Goblet, Saxophone, Pie Pan, Trophy, and Cylinder (see supplemental materials at http://webdocs.cs.ualberta.ca/~yang/conference.htm), are used for testing the proposed approach on real captured data. Here we use an LG IPS monitor to display backdrops and a Point Grey Blackfly monochromatic camera to capture the scene. To automate the capture process, the patterns are displayed at 2fps, while the scene is captured in video mode at 6fps. As a result, three images are captured for each pattern and the middle one is used. This removes the needs for synchronizing between the monitor and the camera.

We first evaluate the effectiveness of CS-based data acquisition by comparing with the conventional DFT method. As shown in Figure 4, our approach achieves comparable results, while requiring only a fraction of sample images. The impact of the region number k is evaluated next. Figure 5 shows that, given the same number of captured images, the matte extraction process is accelerated as the region number k increases. Moreover, as illustrated in Table 1, setting a smaller k value (*e.g.* $k = 2$) requires fewer images while maintains similar visual performance. Hence, the tests on real data further confirms that k offers a tradeoff between the data acquisition process and matte extraction.

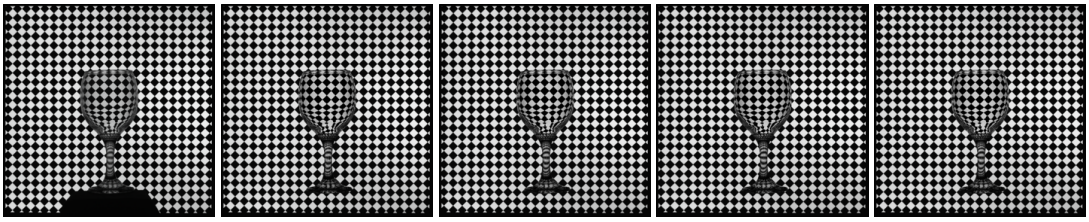
Table 1 compares our approach with the spatial domain method in [7]. It shows that the proposed approach produces more realistic composited results while consuming only a fraction of processing time regardless of the setting of the region number. In the last scene, a goblet is laid on a glossy pie pan. The former object is highly refractive, whereas the latter reflects lights from a fairly broad area of the background. As a result, the contrast and sharpness of the two zoomed-in areas are noticeably different. Our approach properly handles both areas, whereas Duan *et al.* gives blurry output due to the block assumption.

Table 2 further highlights the features of different environmental matting approaches. Although the state-of-the-



(a) Ground Truth (b) Duan *et al.* (c) Ours without Phase (d) Ours with Phase (e) Comparison

Figure 3. Comparison using synthetic data with ground truth (a). The result of [7] (b) is computed by capturing 40 images in the coarse level and 300 images in the fine level. It is blocky because of their square block assumption. Our CS-based approach (c & d) uses 320 images and shows better performance in terms of both accuracy and matte extraction time. The red and blue boxes show zoom in views.



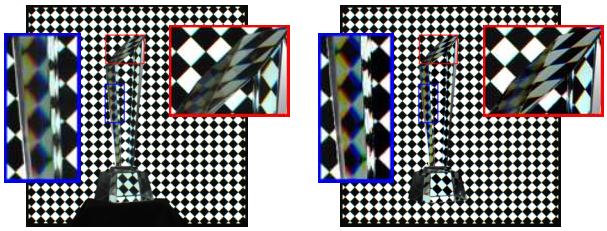
(a) Photograph (b) 200 images by CS (c) 400 images by CS (d) 800 images by CS (e) 2120 images by DFT

Figure 4. CS-based data acquisition on the Goblet object (under phase region $k = 2$). As the number of images increases, the result of CS-based approach improves. With 400 images, the result is visually comparable to the conventional DFT, which requires 2120 images.

art methods [5, 13] can handle multiple-region mapping and produce high quality visual effects, they require thousands of images. In addition, the time-consuming non-linear optimization in [5] depends on a number of parameters that can greatly affect the quality of mattes, while the adaptive data acquisition process in [13] takes hours and requires synchronization between the monitor and the camera. These limit their practical applications. For approaches with low data acquisition requirement, they require block assumption [7, 20] and thus cannot obtain visually pleasing results.

Our approach locates contributing sources of the background at the pixel level and enjoys the following features: 1) Fast data acquisition and matte extraction process; 2) No camera/monitor synchronization or calibration needed; 3) Easy reproducibility with only two parameters, both of which are fixed in our experiments. These make our approach easy to use and can greatly facilitate follow-up applications, *e.g.* 3D reconstruction [11].

Note that we choose to use monochromatic camera in our experiment because the artifacts of Bayer mosaic can be eliminated. This helps to extract wavelength-dependent mattes, resulting proper handling of dispersion effects. As shown in Figure 6, by displaying patterns of different prime colors and performing environment matte extraction separately, we can render the dispersion effect of the object.



(a) Photograph (b) Composite

Figure 6. Handling dispersion. By processing the different color channels separately, our approach is able to render the rainbow phenomenon of the trophy. The greenish fringe around the checkerboard is due to chromatic lens aberration, which is not corrected in our experiment.

6. Conclusions

In this paper, we propose a novel frequency-based environment matting approach, which mainly addresses two major limitations of existing approaches. First, by leveraging CS, we simplify the data acquisition process of the conventional frequency-based environment matting [19]. Second, by augmenting with phase information, we further reduce the measurement cost and accelerate the expensive signal reconstruction process in CS, while accurately locating the contributing sources at the pixel level.

Nevertheless, one limitation of [19] remains unsolved in our approach. That is, for acceleration purpose, we both assume that the unknown light transportation matrix $\mathbf{W}$ can be decomposed into the element-wise product of a row vector and a column vector, *i.e.* Eq.(4). While this assumption has limited impacts on the algorithm's capability in handling contributions from broad areas of the background (*e.g.* "Pie Pan") or contributions from a large number of scat-

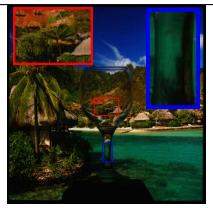
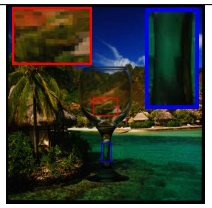
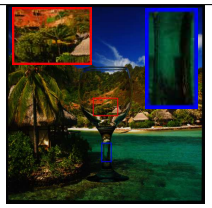
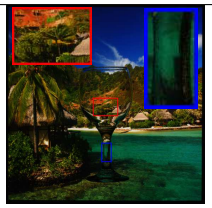
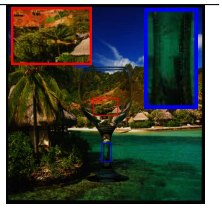
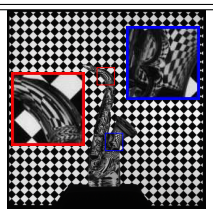
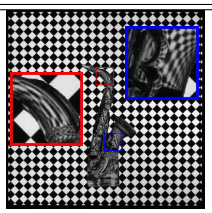
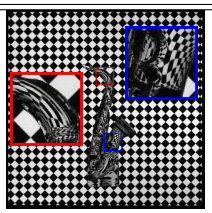
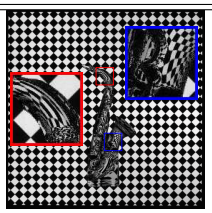
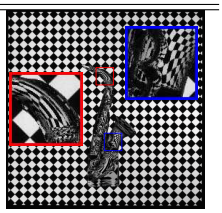
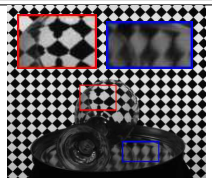
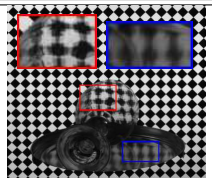
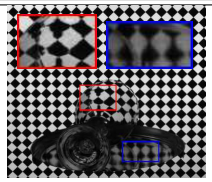
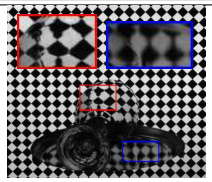
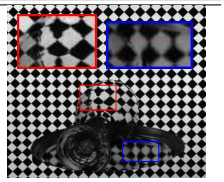
	# Img. (time)	Photograph	Duan *et al.* 340 (2.8)	$k = 2$ 400 (3.3)	$k = 4$ 600 (5)	$k = 8$ 800 (6.7)
Goblet	Composite					
	Runtime		128.6	19.1	16.4	14.4
Saxophone	Composite					
	Runtime		145.5	15.4	14.1	14.4
Pie Pan	Composite					
	Runtime		212.7	24.0	21.5	20.4

Table 1. Comparison with the method in [7] on real data. Note that capturing more images will not improve the results of [7] due to the hierarchical sampling scheme being used. Our results are more visually appealing, while consuming less processing time. Since the scenes are captured at 2fps in the video mode, the number of minutes needed for acquisition are computed as $\#imgs/120$.

Methods	# images when $n = 1024$	Runtime ($n = 1024$)	Remarks
Zongker *et al.* [20]	$\mathcal{O}(\log n)$, 20 images	20 mins when $n = 512$	Single-region mapping, block assumption
Chuang *et al.* [5]	$\mathcal{O}(n)$, 1800 images	Not available	Multi-region mapping, complex optimization
Real Time *et al.* [5]	1 image	2 mins	One-pixel mapping, colorless & pure specular object
Wavelet [13]	$\mathcal{O}(n)$, 2400 images	12 hours	Multi-region mapping, adaptive acquisition
Frequency [19]	$\mathcal{O}(n)$, 4096 images	5-10 mins	Multi-pixel mapping, slow acquisition
Duan *et al.* [7, 8]	$\mathcal{O}(s \log(n^2/s))$, 340 images	See Table 1	Multi-region mapping, block assumption
Ours	$\mathcal{O}(s \log(2n/s))$, 400 images	See Table 1	Multi-pixel mapping, fast acquisition & extraction

Table 2. Comparisons among different environment matting methods, where s denotes the sparsity of a signal. The information about the previous methods is directly copied or estimated from the corresponding papers. Note that the extraction time of [13] includes data acquisition because of its adaptive scheme.

tered sources, it may introduce visual artifacts when a foreground pixel has two non-adjacent dominating contributing regions. For example, the foreground highlighted by the red box in Figure 6 mainly receives lights from two regions in the background, one coming from refraction and the other from reflection. Under the above assumption, our approach may either locate additional but incorrect contributing sources or lose the weaker one. Thus the composited result may appear different from the photograph. In the future, we plan to address this ambiguity problem using additional diagonal patterns [5].

The environment matting problem is addressed in this paper using the proposed CS-based framework in the frequency domain. We argue that our proposed framework is also applicable to other many-to-one decomposition problems, *e.g.* dual photograph [15], where many projector pixels are merged into one camera pixel. CS has been utilized to reduce the complexity of data acquisition in dual photography [16], whereas the process of reconstructing the light reflection functions is very slow (almost 3 hours on a 24-node cluster for rendering a 256×256 image). Hence, we plan to apply our new frequency-based framework to tackle the dual photography problem in the near future.

Acknowledgements. We thank NSERC and the University of Alberta for the financial support, and the anonymous reviewers for their constructive comments.

References

[1] Persistence of vision (tm) raytracer. http://www.povray.org/. 5

[2] E. J. Candes. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1433–1452, 2006. 2

[3] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006. 2

[4] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *Computer Vision–ECCV 2008*, pages 155–168. Springer, 2008. 2

[5] Y.-Y. Chuang, D. E. Zongker, J. Hindorff, B. Curless, D. H. Salesin, and R. Szeliski. Environment matting extensions: Towards higher accuracy and real-time capture. In *Proceedings of ACM SIGGRAPH 00*, pages 121–130. ACM Press/Addison-Wesley Publishing Co., 2000. 1, 4, 7, 8

[6] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006. 2

[7] Q. Duan, J. Cai, and J. Zheng. Compressive environment matting. *The Visual Computer*, pages 1–14, 2014. 1, 2, 3, 4, 6, 7, 8

[8] Q. Duan, J. Cai, J. Zheng, and W. Lin. Fast environment matting extraction using compressive sensing. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011. 2, 8

[9] J. Huang, X. Huang, and D. Metaxas. Learning with dynamic group sparsity. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 64–71. IEEE, 2009. 2, 4, 5

[10] N. Hurley and S. Rickard. Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741, 2009. 3

[11] K. N. Kutulakos and E. Steger. A theory of refractive and specular 3d shape by light-path triangulation. *International Journal of Computer Vision*, 76(1):13–29, 2008. 7

[12] C. L. Lawson and R. J. Hanson. *Solving least squares problems*, volume 161. SIAM, 1974. 5

[13] P. Peers and P. Dutré. Wavelet environment matting. In *Proceedings of the 14th Eurographics workshop on Rendering*, pages 157–166. Eurographics Association, 2003. 1, 2, 7, 8

[14] P. Peers, D. K. Mahajan, B. Lamond, A. Ghosh, W. Matusik, R. Ramamoorthi, and P. Debevec. Compressive light transport sensing. *ACM Transactions on Graphics (TOG)*, 28(1):3, 2009. 2

[15] P. Sen, B. Chen, G. Garg, S. R. Marschner, M. Horowitz, M. Levoy, and H. Lensch. Dual photography. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 745–755. ACM, 2005. 8

[16] P. Sen and S. Darabi. Compressive dual photography. In *Computer Graphics Forum*, volume 28, pages 609–618. Wiley Online Library, 2009. 2, 8

[17] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman. Image-based environment matting. In *Rendering Techniques*, pages 279–290, 2002. 1, 4

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009. 2

[19] J. Zhu and Y.-H. Yang. Frequency-based environment matting. In *Computer Graphics and Applications, 2004. PG 2004. Proceedings. 12th Pacific Conference on*, pages 402–410. IEEE, 2004. 1, 2, 3, 5, 7, 8

[20] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin. Environment matting and compositing. In *Proceedings of ACM SIGGRAPH 99*, pages 205–214. ACM Press / ACM SIGGRAPH / Addison Wesley Logman, July 1999. 1, 2, 7, 8