Image Signature: Highlighting Sparse Salient Regions

Xiaodi Hou, Jonathan Harel, and Christof Koch, *Member*, *IEEE*

Abstract—We introduce a simple image descriptor referred to as the *image signature*. We show, within the theoretical framework of sparse signal mixing, that this quantity spatially approximates the foreground of an image. We experimentally investigate whether this approximate foreground overlaps with visually conspicuous image locations by developing a saliency algorithm based on the image signature. This saliency algorithm predicts human fixation points best among competitors on the Bruce and Tsotsos [1] benchmark data set and does so in much shorter running time. In a related experiment, we demonstrate with a *change blindness* data set that the distance between images induced by the image signature is closer to human perceptual distance than can be achieved using other saliency algorithms, pixel-wise, or GIST [2] descriptor methods.

Index Terms—Saliency, visual attention, change blindness, sign function, sparse signal analysis.

1 Introduction

THE problem of finding all objects in a scene and separating them from the background is known as *figure-ground separation*. The brain can perform this separation very quickly [3], and doing so on a machine remains a major challenge for engineers and scientists. The problem is closely related to many of the core applications of machine vision, including scene understanding, content-based image retrieval, object recognition, and tracking. In this paper, we provide an approach to the figure-ground separation problem using a binary, holistic image descriptor called the "image signature." It is defined as the sign function of the Discrete Cosine Transform (DCT) of an image. As we shall demonstrate, this simple descriptor preferentially contains information about the foreground of an image—a property which we believe underlies the usefulness of this descriptor for detecting salient image regions.

In Section 2, we formulate the figure-ground separation problem in the framework of sparse signal analysis. We prove that the Inverse Discrete Cosine Transform (IDCT) of the image signature concentrates the image energy at the locations of a spatially sparse foreground, relative to a spectrally sparse background. Then, in Section 3.1, we demonstrate this phenomenon on synthetic images with sparse foregrounds much weaker in intensity than the complex background pattern.

Two experiments are presented to quantify the relationship between the image signature and human visual attention. In Section 3.2, we demonstrate that a saliency map derived from the image signature outperforms many leading saliency algorithms on a benchmark data set of eye-movement fixation points. In Section 3.3, we introduce reaction time data collected from nine subjects in a change blindness experiment. We show that the distance between images induced by the image signature most closely matches the perceptual distance between images inferred from these data among competing measures derived from other saliency algorithms, the GIST descriptor, and simpler pixel measures.

- *X. Hou is with the Department of Computation and Neural Systems, MC 216-76, California Institute of Technology, Pasadena, CA 91125. E-mail: xhou@caltech.edu.*
- *J. Harel is with the Department of Electrical Engineering, MC 216-76, California Institute of Technology, Pasadena, CA 91125. E-mail: harel@klab.caltech.edu.*
- *C. Koch is with the Department of Biology and Computation and Neural Systems, California Institute of Technology, MC 216-76, Pasadena, CA 91125, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea. E-mail: koch@klab.caltech.edu.*

Manuscript received 7 Jan. 2011; revised 21 Apr. 2011; accepted 1 June 2011; published online 19 July 2011.
Recommended for acceptance by S. Avidan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-01-0014.
Digital Object Identifier no. 10.1109/TPAMI.2011.146.

1.1 Related Work

Holistic image processing short-circuits the need for segmentation, key-point matching, and other local operations. Bolstered by growing general interest in large-scale image retrieval systems, holistic image descriptors have become a topic of intense study in the computer vision literature. GIST [2] is an excellent example of such an algorithm in this field. Other holistic scene models focus on the separation of foreground and background. For example, Candes et al. [4] introduced a sparse matrix factorization model.

A more relevant study comes from Hou and Zhang [5], motivated by Oppenheim et al.'s early discovery [6], [7]. They found that the residual Fourier amplitude spectrum, the difference between the original Fourier amplitude spectrum and its smoothed copy, could be used to form a saliency map. The residual retains more high-frequency information than low, where the smoothed copy is similar to the original. The image signature, in comparison, discards amplitude information across the entire frequency spectrum, storing only the sign of each DCT component, equivalent to phase for a Fourier decomposition. The image signature is thus very compact, with a single bit per component, and as we shall show in the remainder of this paper, possesses important properties related to the foreground of an image.

2 Image Signature

2.1 Preliminaries

We begin by considering gray-scale images which exhibit the following structure:

$$\mathbf{x} = \mathbf{f} + \mathbf{b}, \quad \mathbf{x}, \mathbf{f}, \mathbf{b} \in \mathbb{R}^N, \tag{1}$$

where $\mathbf{f}$ represents the foreground or figure signal and is assumed to be sparsely supported in the standard spatial basis. $\mathbf{b}$ represents the background and is assumed to be sparsely supported in the basis of the Discrete Cosine Transform. In other words, both $\mathbf{f}$ and $\hat{\mathbf{b}}$ have only a small number of nonzero components. Please refer to Table 1 for important definitions used throughout the rest of this section.

Performing the exact separation between $\mathbf{b}$ and $\mathbf{f}$ given only $\mathbf{x}$ and the fact of their sparseness is, in general, very difficult. For the problem of figure-ground separation, we are only interested in the spatial support of $\mathbf{f}$ (the set of pixels for which $\mathbf{f}$ is nonzero). In this paper, we show, first analytically, then empirically, that given an image which can be decomposed as 1, we can approximately isolate the support of $\mathbf{f}$ by taking the sign of the mixture signal $\mathbf{x}$ in the transformed domain and then inversely transform it back into the spatial domain, i.e., by computing the reconstructed image $\bar{\mathbf{x}} = \mathrm{IDCT}[\mathrm{sign}(\hat{\mathbf{x}})]$. Formally, the image signature is defined as

$$\mathrm{ImageSignature}(\mathbf{x}) = \mathrm{sign}(\mathrm{DCT}(\mathbf{x})). \tag{2}$$

If we assume that an image foreground is visually conspicuous relative to its background, then we can form a *saliency map* $\mathbf{m}$ (see [8] for classic use) by smoothing the squared reconstructed image defined above

$$\mathbf{m} = g * (\bar{\mathbf{x}} \circ \bar{\mathbf{x}}), \tag{3}$$

TABLE 1
Important Notation and Terms Used in This Paper

$\hat{\mathbf{x}}$	$\mathrm{DCT}(\mathbf{x})$.
$\mathrm{sign}(\mathbf{x})$	The entrywise sign operator.
$\bar{\mathbf{x}}$	$\mathrm{IDCT}[\mathrm{sign}(\hat{\mathbf{x}})]$, the reconstructed image.
$T_{\mathbf{x}}$	Support set of $\mathbf{x}$.
$\Omega_{\mathbf{x}}$	Support set of $\hat{\mathbf{x}}$.
$\lvert x\rvert$	The absolute value of a real number x.
$\lvert\mathcal{S}\rvert$	The cardinality of a set $\mathcal{S}$.
$\lVert\mathbf{x}\rVert_p$	The ℓ^p norm of vector $\mathbf{x}$ ($p=2$ if omitted).
$\langle\mathbf{x},\mathbf{y}\rangle$	The inner-product of $\mathbf{x}$ and $\mathbf{y}$.
$E(X)$	The expectation of random variable X.
$\circ$	The Hadamard (entrywise) product operator.
$*$	The convolution operator.

where g is a Gaussian kernel. Our experiments in Section 3.1 show that a simple Gaussian smoothing is necessary here because the support $T_{\mathbf{f}}$ of a salient object is usually not only spatially sparse, but also localized in a contiguous region.

We also define a distance metric D between images $\mathbf{x}^1$ and $\mathbf{x}^2$ based on the ℓ^0 distance between image signatures (viz., the Hamming distance):

$$D(\mathbf{x}^1,\mathbf{x}^2)=\|\mathrm{sign}(\hat{\mathbf{x}}^1)-\mathrm{sign}(\hat{\mathbf{x}}^2)\|_0. \tag{4}$$

Building on the idea that the image signature preferentially contains foreground information, this subtraction compares the sparse foreground information in two images, without explicitly first computing either $\mathbf{b}$ or $\mathbf{f}$. Later, we provide empirical evidence for the utility of this metric.

2.2 Image Signature: Foreground Properties

In this section, we provide evidence that, for an image which adheres to a certain mathematical structure, the image signature can be used to approximately obtain the location of the foreground.

Proposition 1 (Signature suppresses background). *The image reconstructed from the image signature approximates the location of a sufficiently sparse foreground on a sufficiently sparse background as follows:*

$$E\left(\frac{\langle\bar{\mathbf{f}},\bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{f}}\|\cdot\|\bar{\mathbf{x}}\|}\right)\geq 0.5,\quad \text{for } |\Omega_{\mathbf{b}}|<\frac{N}{6}. \tag{5}$$

Proof. Our proof is based on the Uniform Uncertainty Principle (UUP) proposed by Candes and Tao [9]. Let Θ be a subset of $\{1,2,\ldots,N\}$ of size $|\Theta|$. UUP states that if $\mathbf{f}$ is sufficiently spatially sparse, that is, if

$$|T_{\mathbf{f}}|\leq \alpha|\Theta|/\lambda, \tag{6}$$

where λ is the oversampling factor and α is a sufficiently small constant, then with an overwhelming probability, the energy of $\hat{\mathbf{f}}$ supported on Θ is bounded:

$$\frac{|\Theta|}{2N}\|\mathbf{f}\|\leq\|\hat{\mathbf{f}}\circ\mathbf{1}_\Theta\|\leq 3\frac{|\Theta|}{2N}\|\mathbf{f}\|, \tag{7}$$

where $\mathbf{1}_\Theta$ is the vector with zeros at component indices not in Θ and ones at component indices in Θ.

The oversampling factor depends on the choice of transform. Rudelson and Vershynin [10] show that for Fourier transform, $\lambda=O(\log^5 N)$. Because of the similarities between DCT and DFT, and that images are real valued, this factor is the same for the DCT. In fact, one can construct a signal $\mathbf{x}'\in\mathbb{R}^{4N}$ from the original $\mathbf{x}\in\mathbb{R}^N$ as the following:

$$\begin{aligned} x'_{2n}&=x_n & x'_{2n-1}&=0\\ x'_{4N-2n+2}&=x_{N-n+1} & x'_{4N-2n+1}&=0,\end{aligned}$$

such that $\mathrm{DFT}(\mathbf{x})$ exactly equals $\mathrm{DCT}(\mathbf{x}')$.

According to Plancherel's theorem, we have $\|\mathbf{f}\|=\|\hat{\mathbf{f}}\circ\mathbf{1}_{\Omega_{\mathbf{f}}}\|$. Then, the following inequality can be derived from UUP (7):

$$\begin{aligned} 3\frac{|\Omega_{\mathbf{f}}|}{2N}\|\mathbf{f}\|&\geq\|\hat{\mathbf{f}}\circ\mathbf{1}_{\Omega_{\mathbf{f}}}\|\\ 3\frac{|\Omega_{\mathbf{f}}|}{2N}&\geq 1\\ |\Omega_{\mathbf{f}}|&\geq\frac{2}{3}N.\end{aligned} \tag{8}$$

Recall that (8) holds with overwhelming probability only if $\mathbf{f}$ the foreground is sufficiently spatially sparse in the sense of (6).

From this, we estimate $\langle\bar{\mathbf{f}},\bar{\mathbf{x}}\rangle$:

$$\begin{aligned}\langle\bar{\mathbf{f}},\bar{\mathbf{x}}\rangle&=\langle\mathrm{IDCT}[\mathrm{sign}(\hat{\mathbf{f}})],\mathrm{IDCT}[\mathrm{sign}(\hat{\mathbf{x}})]\rangle\\&=\mathrm{IDCT}[\langle\mathrm{sign}(\hat{\mathbf{f}}),\mathrm{sign}(\hat{\mathbf{x}})\rangle]\\&=\langle\mathrm{sign}(\hat{\mathbf{f}}),\mathrm{sign}(\hat{\mathbf{x}})\rangle\\&=\sum_{i\in\Omega_{\mathbf{b}}}\mathrm{sign}(\hat{f}_i)\cdot\mathrm{sign}(\hat{x}_i)\\&\quad+\sum_{j\notin\Omega_{\mathbf{b}}}\mathrm{sign}(\hat{f}_j)\cdot\mathrm{sign}(\hat{x}_j).\end{aligned} \tag{9}$$

Since $\mathbf{f}$ and $\mathbf{b}$ are independent of each other, we assume

$$\mathbb{P}(\mathrm{sign}(\hat{f}_i)=\mathrm{sign}(\hat{b}_i)|i\in\Omega_{\mathbf{b}})=0.5,$$

where $\mathbb{P}$ stands for probability. Then,

$$\begin{aligned}\mathbb{P}(\mathrm{sign}(\hat{f}_i)=\mathrm{sign}(\hat{x}_i)|i\in\Omega_{\mathbf{b}})&=\mathbb{P}(\mathrm{sign}(\hat{f}_i)=\mathrm{sign}(\hat{b}_i)|i\in\Omega_{\mathbf{b}})\\&\quad+\mathbb{P}(|\hat{f}_i|>|\hat{b}_i|,\mathrm{sign}(\hat{f}_i)\neq\mathrm{sign}(\hat{b}_i)|i\in\Omega_{\mathbf{b}})\geq 0.5.\end{aligned}$$

Therefore,

$$E\left[\sum_{i\in\Omega_{\mathbf{b}}}\mathrm{sign}(\hat{f}_i)\cdot\mathrm{sign}(\hat{x}_i)\right]\geq 0. \tag{10}$$

Since $\mathrm{sign}(\hat{b}_j)=0$, for $j\notin\Omega_{\mathbf{b}}$, we have

$$\sum_{j\notin\Omega_{\mathbf{b}}}\mathrm{sign}(\hat{f}_j)\cdot\mathrm{sign}(\hat{x}_j)=\sum_{j\notin\Omega_{\mathbf{b}}}\mathrm{sign}(\hat{f}_j)^2\geq|\Omega_{\mathbf{f}}|-|\Omega_{\mathbf{b}}|. \tag{11}$$

Combining (10), (11), and (9), we have

$$E\left(\frac{\langle\bar{\mathbf{f}},\bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{f}}\|\cdot\|\bar{\mathbf{x}}\|}\right)\geq\frac{|\Omega_{\mathbf{f}}|-|\Omega_{\mathbf{b}}|}{\sqrt{|\Omega_{\mathbf{f}}|\cdot|\Omega_{\mathbf{x}}|}}\geq\frac{|\Omega_{\mathbf{f}}|-|\Omega_{\mathbf{b}}|}{N}. \tag{12}$$

Given the bound provided by (8),

$$E\left(\frac{\langle\bar{\mathbf{f}},\bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{f}}\|\cdot\|\bar{\mathbf{x}}\|}\right)\geq\frac{2}{3}-\frac{1}{N}|\Omega_{\mathbf{b}}|\geq 0.5,$$

if we assume that the background $\mathbf{b}$ is sufficiently sparse $|\Omega_{\mathbf{b}}|<N/6$. □

An important note is that Proposition 1 does not depend on the relative energies of the foreground and background, $\|\mathbf{f}\|$ and $\|\mathbf{b}\|$, only their sparseness. This will later be demonstrated empirically in Section 3.1, Fig. 3, on synthetic data. Proposition 1 does not establish a direct relationship between the reconstructed image $\bar{\mathbf{x}}$ and the actual foreground $\mathbf{f}$; instead the relationship is to a function of the foreground, $\bar{\mathbf{f}}$. We will now show that $\bar{\mathbf{f}}$ contains important information about the spatial support of the foreground, $T_{\mathbf{f}}$.

Proposition 2. *For a foreground signal* $\mathbf{f}$ *with nonzero elements independently drawn from the unit Gaussian distribution, over 79 percent of* $\bar{\mathbf{f}}$ *is expected to be contained in the support of the foreground* $T_{\mathbf{f}}$*. Namely,*

$$E\left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}\right) \geq \sqrt{\frac{2}{\pi}} \approx 0.7979 \text{ where}$$
$$\alpha = \sqrt{\sum_{i\in T_{\mathbf{f}}} \bar{f}_i^2}, \text{ and}$$
$$\beta = \sqrt{\sum_{j\notin T_{\mathbf{f}}} \bar{f}_j^2}.$$

Proof. First, we quantify the expected correlation between $\hat{\mathbf{f}}$ and $\text{sign}(\hat{\mathbf{f}})$:

$$E\left(\frac{\langle \hat{\mathbf{f}}, \text{sign}(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\| \cdot \|\text{sign}(\hat{\mathbf{f}})\|}\right) = E\left(\frac{\sum_i \hat{f}_i \cdot \text{sign}(\hat{f}_i)}{\|\hat{\mathbf{f}}\| \cdot \|\text{sign}(\hat{\mathbf{f}})\|}\right) = E\left(\frac{\sum_i |\hat{f}_i|}{\|\hat{\mathbf{f}}\| \cdot \|\text{sign}(\hat{\mathbf{f}})\|}\right). \tag{13}$$

For zero-mean unit-variance normally distributed f_i,

$$E\left(\frac{\langle \hat{\mathbf{f}}, \text{sign}(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\| \cdot \|\text{sign}(\hat{\mathbf{f}})\|}\right) = E(|\hat{f}_i|) = \sqrt{\frac{2}{\pi}}. \tag{14}$$

Then, we show that the amount of energy of $\bar{\mathbf{f}}$ that falls into $T_{\mathbf{f}}$ is lower bounded.

Because the correlation between a pair of signals in the spatial domain equals their correlation in the DCT domain, we have

$$\frac{\langle \hat{\mathbf{f}}, \text{sign}(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\| \cdot \|\text{sign}(\hat{\mathbf{f}})\|} = \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\| \cdot \|\bar{\mathbf{f}}\|} = \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\| \sqrt{\sum_{i\in T_{\mathbf{f}}} \bar{f}_i^2 + \sum_{j\notin T_{\mathbf{f}}} \bar{f}_j^2}} = \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\| \sqrt{\alpha^2+\beta^2}}. \tag{15}$$

Let $\mathbf{1}_{T_{\mathbf{f}}}$ be the indicator function that has the value 1 for all elements of $T_{\mathbf{f}}$ and 0 elsewhere. From the Cauchy-Schwartz inequality,

$$\langle \mathbf{f}, \bar{\mathbf{f}}\rangle = \langle \mathbf{f} \circ \mathbf{1}_{T_{\mathbf{f}}}, \bar{\mathbf{f}} \circ \mathbf{1}_{T_{\mathbf{f}}}\rangle \leq \alpha \|\mathbf{f}\|. \tag{16}$$

According to (14) and (16),

$$E\left(\frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\| \cdot \|\bar{\mathbf{f}}\|}\right) = \sqrt{\frac{2}{\pi}} \leq E\left(\frac{\alpha\|\mathbf{f}\|}{\|\mathbf{f}\| \cdot \sqrt{\alpha^2+\beta^2}}\right)$$
$$E\left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}\right) \geq \sqrt{\frac{2}{\pi}} \approx 0.7979. \tag{17}$$

□

2.3 Hamming Distance Captures the Angular Difference Between Images Which Share a Background

As we have suggested in (4), the Hamming distance D between two image signatures can be used as a distance metric. We show below how this distance is related to the angular difference between a pair of images $\mathbf{x}^1$ and $\mathbf{x}^2$.

Let ϕ_i denote the ith basis function of the DCT. We assume that ϕ_i is independent of both $\mathbf{x}^1$ and $\mathbf{x}^2$. From [11] (Lemma 3.2), we know that

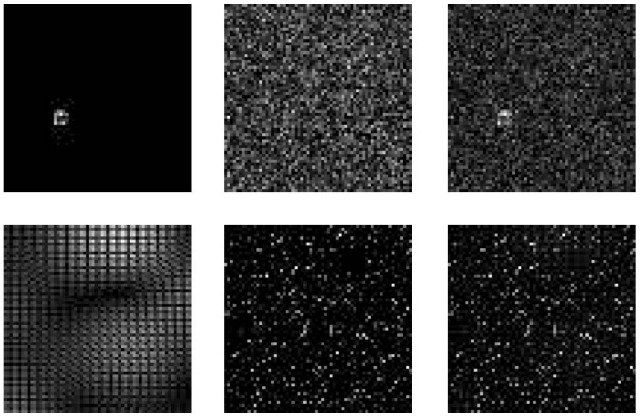

Fig. 1. An illustration of the randomly generated images. The first row: $\mathbf{f}$, $\mathbf{b}$, and $\mathbf{x}$ in the spatial domain. The second row: The same signals represented in the DCT domain: $\hat{\mathbf{f}}$, $\hat{\mathbf{b}}$, and $\hat{\mathbf{x}}$.

$$\mathbb{P}[\text{sign}(\langle \mathbf{x}^1, \phi_i\rangle) \neq \text{sign}(\langle \mathbf{x}^2, \phi_i\rangle)] = \frac{1}{\pi} cos^{-1}\left(\frac{\langle \mathbf{x}^1, \mathbf{x}^2\rangle}{\|\mathbf{x}^1\| \cdot \|\mathbf{x}^2\|}\right).$$

Let d_i be the indicator function

$$d_i = \begin{cases} 0 & \text{if } \text{sign}(\langle \mathbf{x}^1, \phi_i\rangle) = \text{sign}(\langle \mathbf{x}^2, \phi_i\rangle) \\ 1 & \text{if } \text{sign}(\langle \mathbf{x}^1, \phi_i\rangle) \neq \text{sign}(\langle \mathbf{x}^2, \phi_i\rangle), \end{cases}$$

and $D = \|\text{sign}(\hat{\mathbf{x}}^1) - \text{sign}(\hat{\mathbf{x}}^2)\|_0 = \sum_i^N d_i$ since $\langle \mathbf{x}, \phi_i\rangle = \hat{x}_i$. Then, the Chernoff bounds guarantee that

$$\forall \epsilon > 0, \quad \mathbb{P}\big(D > (1+\epsilon)N\mu\big) < e^{-\frac{1}{4}N\mu\epsilon^2}$$
$$\forall 0 < \epsilon < 1, \quad \mathbb{P}\big(D < (1-\epsilon)N\mu\big) < e^{-\frac{1}{2}N\mu\epsilon^2},$$

where $\mu = E(d_i)$. This result indicates that for large enough N, the following statement is true with high probability:

$$(1-\epsilon)\frac{D}{N} \leq \frac{1}{\pi} cos^{-1}\left(\frac{\langle \mathbf{x}^1, \mathbf{x}^2\rangle}{\|\mathbf{x}^1\| \cdot \|\mathbf{x}^2\|}\right) \leq (1+\epsilon)\frac{D}{N}. \tag{18}$$

Suppose that the pair of images $\mathbf{x}^1$ and $\mathbf{x}^2$ share the same background, i.e., $\mathbf{x}^i = \mathbf{b} + \mathbf{f}^i$. Then, for a spatially sparse foreground, the images will be different in only a few pixels and the distance D will be much less than $N/2$, and very sensitive to the difference in the pixels in the foreground. However, if the two images do not share a background, then most of their pixels will be independent and the quantity in (18) will be very close to 0.5 and insensitive to small changes in foreground pixels.

3 The Experiments

3.1 Image Signature on Synthetic Images

In the previous section, we provided theoretical arguments connecting the image signature to the spatial support of a sparse foreground. In this section, we use synthetic images to demonstrate its behavior in carefully constructed cases. In later sections, we will demonstrate the utility of the image signature for practical applications.

Let $\mathbf{f}, \mathbf{b}, \mathbf{x} \in \mathbb{R}^{64\times 64}$. The support of the foreground is a 5×5 block ($|T_{\mathbf{f}}| = 25$) that appears at a random location. The support for $\hat{\mathbf{b}}$ is randomly selected in the DCT domain, with $|\Omega_{\mathbf{b}}| = 500$. For $i \in T_{\mathbf{f}}$, the amplitude of each pixel f_i is drawn from a normal distribution. Similarly, for $j \in \Omega_{\mathbf{b}}$, each $\hat{b}_j$ is drawn from normal distribution. Fig. 1 shows $\mathbf{f}$, $\mathbf{b}$, and $\mathbf{x}$ in both the spatial and the DCT domains.

The image signature reconstruction is illustrated in Fig. 2. Note that a Gaussian blurring is used to suppress the noise introduced by the sign quantization. Ideally, the standard deviation σ of the Gaussian kernel should be proportional to the size of the object of

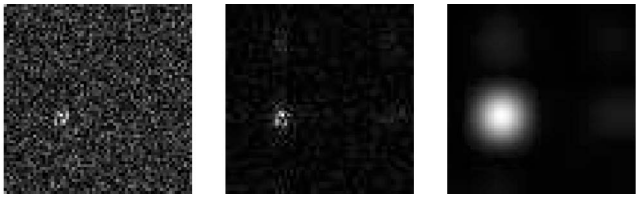

Fig. 2. An example of the input image $\mathbf{x}$, the reconstructed image $\bar{\mathbf{x}}$, and the saliency map $\mathbf{m}$.

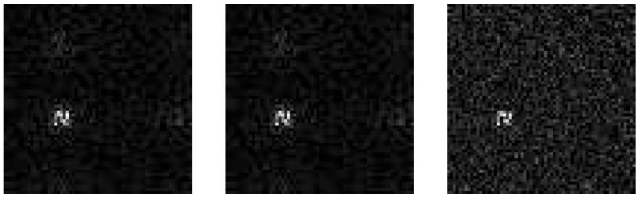

Fig. 3. The reconstructed $\bar{\mathbf{x}}$ with foreground reweighted as 10^{-5}, 10^{-10}, and 10^{-15}. The saliency maps of these signals remain almost the same despite a huge difference in the foreground amplitude.

interest. We here choose $\sigma = 0.05$ of the image width (in other words, we implicitly assume that the width of the object is about 10 percent of the image width).

From Proposition 1, it follows that the reconstructed image $\bar{\mathbf{x}}$ should not be very affected by the amplitude of the foreground, instead only its spatial support. We tested this by multiplying the amplitude of the foreground $\mathbf{f}$ by a factor of 10^{-5}, 10^{-10}, and 10^{-15}, while holding the background completely constant. Surprisingly, the reconstructed signal $\bar{\mathbf{x}}$ is not changed by the foreground energy until it approaches a minimal numerical value in Matlab, 2.2×10^{-16}.

Proposition 2 guarantees that the majority of the foreground energy stays in the support of the foreground after the sign quantization. Our theoretical justification is based on a Gaussian distribution. However, it has been suggested by Ruderman [12] that the histogram of pixel intensity of natural images follows a power law (that is, the pixel intensity follows a Pareto distribution). We generated the foreground pixels based on three different distributions (normal distribution, uniform distribution, and Pareto distribution with the PDF $f(x) = (1+x)^{-2}$), and tested whether the energy of $\bar{\mathbf{x}}$ was constrained in the foreground region. For fair comparisons, the foreground was normalized to $[0, 1]$. The proportion (in the sense of Proposition 2) that fell into $T_{\mathbf{f}}$ was: 79.8, 75.6, and 79.3 percent for the three distributions, respectively.

In some scenarios, the background may not be ideally sparse. In Fig. 4, we provide an empirical demonstration to test the robustness of this method with respect to nonsparse backgrounds. We observe a clear trend that the energy within $T_{\mathbf{f}}$ drops as the complexity of the background increases. It is worth mentioning that even with fairly complex backgrounds (with $|\Omega_{\mathbf{b}}| = 3{,}000$), the saliency map still clearly shows the shape of the foreground support of the image.

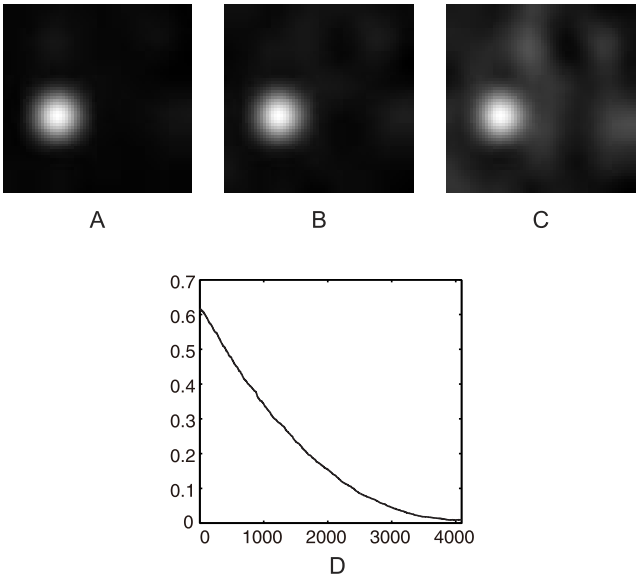

Fig. 4. Performance of the signature algorithm with different background complexity on a 64×64 image ($N = 4{,}096$). A. $|\Omega_{\mathbf{b}}| = 1{,}600$. B. $|\Omega_{\mathbf{b}}| = 2{,}400$. C. $|\Omega_{\mathbf{b}}| = 3{,}200$. D. Proportion of energy concentrated in the foreground support $T_{\mathbf{f}}$, as a function of background cardinality $|\Omega_{\mathbf{b}}|$. The proportion of energy is the square of the fraction provided in Proposition 2.

3.2 Generating the Saliency Map of an Image

Here, we report our experimental findings in saliency detection using the image signature. As we demonstrated earlier, the reconstructed image detects spatially sparse signals embedded in spectrally sparse backgrounds. We will show that the saliency map (3) formed from the reconstruction greatly overlaps with regions of human overt attentional interest, measured as fixation points on an input image.

The exact details of the saliency algorithm are as follows: First, a color image is resized to a coarse 64×48 pixel representation. Then, for each color channel $\mathbf{x}^i$, the saliency map is formed from the image reconstructed from the image signature

$$\mathbf{m} = g * \sum_i (\bar{\mathbf{x}}^i \circ \bar{\mathbf{x}}^i). \tag{19}$$

The standard deviation of the Gaussian blurring kernel g will be discussed in greater detail in the following section.

For the choice of color channels, we use both RGB and CIELAB color spaces. In the following sections, the algorithms associated with these choices will be referred to as **RGB-Signature** and **Lab-Signature**, respectively. An illustration of this RGB-Signature algorithm is shown in Fig. 5.

3.2.1 *Predicting Human Fixation*

To validate the saliency maps generated by our algorithm, we use the data set of human eye-tracking data introduced by Bruce and Tsotsos [1] to compare the various saliency map algorithms

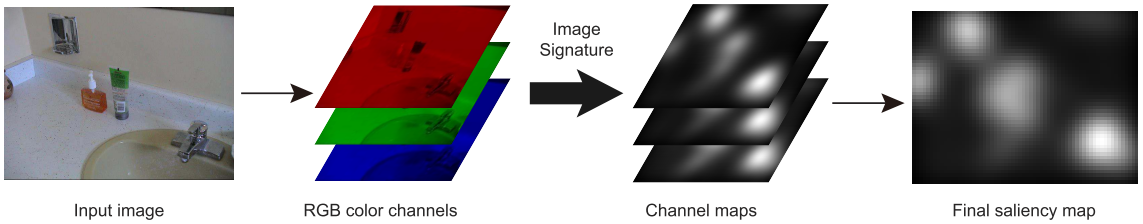

Fig. 5. An illustration of the RGB-Signature algorithm. The input color image is decomposed into three channels. A saliency map is computed for each color channel independently, and the final saliency map is simply the sum across three.

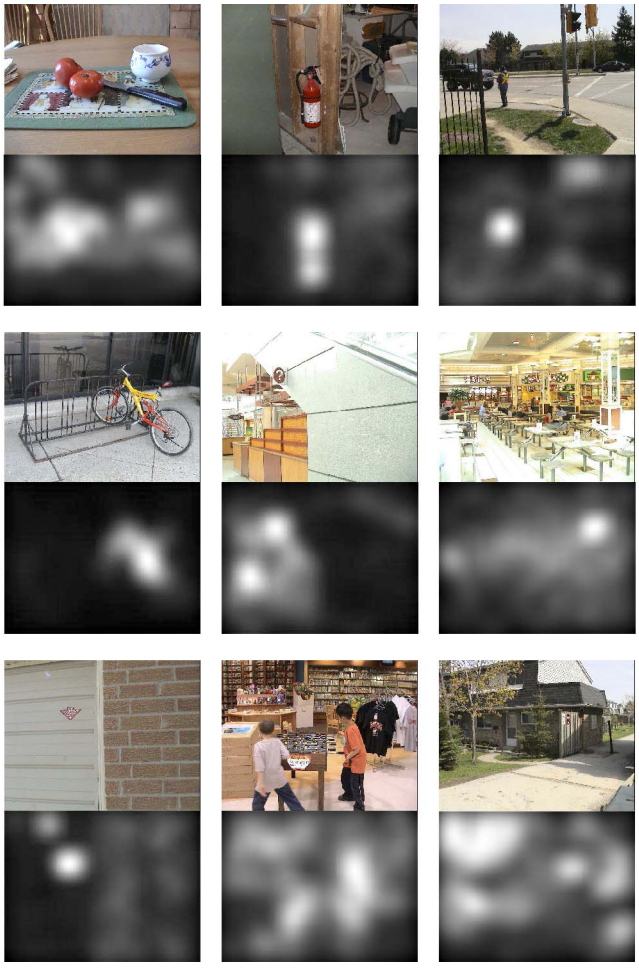

Fig. 6. Sample images from the Bruce data set and their corresponding saliency maps using the Lab-Signature algorithm with $\sigma = 0.05$.

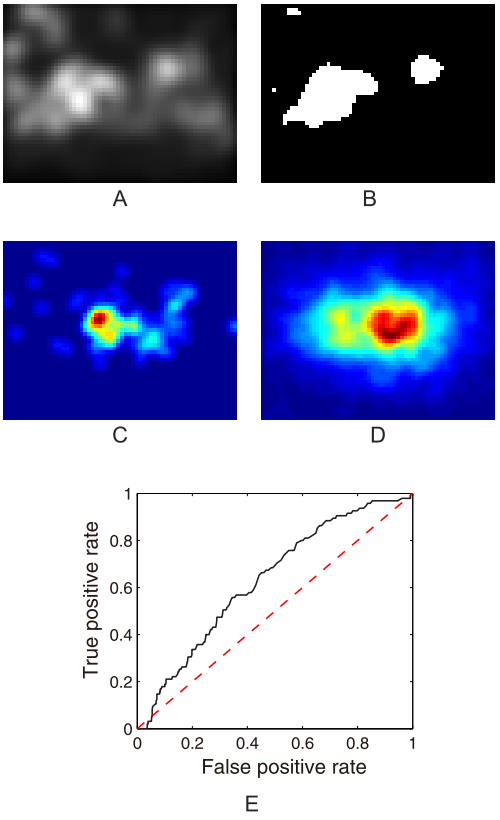

Fig. 7. An illustration of the AUC computation on the first image in Fig. 6. A. The saliency map generated by Lab-Signature algorithm. B. The binary map (thresholded at $T = 0.5$). C. The positive sample set of human fixations on this image (represented as a heat map). D. The negative sample set of human fixations, containing all fixations across the entire data set, except those contained in the positive sample set (represented as a heat map). Both C and D are smoothed for display clarity, but the AUC computation uses the exact fixation points. E. The blue curve shows the ROC curve of Lab-Signature algorithm on this image, with the red reference line indicating the chance level. The area under the blue curve is 0.6329.

here. It consists of 20 subjects free-viewing 120 color images (681×511 pixels) for 4 seconds each. Some sample images are shown in Fig. 6. In order to evaluate the consistency between a particular saliency map and a set of fixations of the image, we computed an ROC Area Under the Curve (AUC) score for each image. As Tatler et al. [13] and Zhang et al. [14] have pointed out, human fixations have strong center-bias which may affect the performance of a saliency algorithm. To remove this center bias, we follow the procedure of Tatler et al. [13]: For one image, the positive sample set is composed of the fixation points of all subjects on that image, whereas the negative sample set is composed of the union of all fixation points across all images from the same data set—except for the positive samples. Each saliency map generated by the algorithm is thresholded and then considered as a binary classifier to separate the positive samples from negative samples. At a particular threshold level T, the true positive rate is the proportion of the positive samples that fall in the positive (white) region of the binary saliency map (Fig. 7B). The false positive rate can be computed in a similar way by using the negative sample set. Sweeping over thresholds yields an ROC curve, of which the area beneath provides a good measure of the power of the saliency map to accurately predict where fixations occurred on an image. Chance level is 0.5, and perfect prediction is 1.0.

We compare our saliency maps generated from the image signature to the following published saliency algorithms: the original Itti-Koch saliency model [8] (denoted **Itti**), Dynamic Visual Attention model [15] (denoted **DVA**), Graph-Based visual saliency [16] (denoted **GBVS**), Attention based on Information maximization [17] (denoted **AIM-original**), and Saliency Using Natural image statistic [14] (denoted **SUN-original**) for comparison. All of the algorithms are based on the original Matlab implementations available on the authors' websites.

An important note about these experiments is that the AUC score is quite sensitive to blurring a saliency map. Some kind of smoothing has been explicitly or implicitly included in most of the algorithms. In order to make a fair comparison, we parameterize the standard deviation of the blurring kernel, and evaluate the performance of an algorithm under different blurring conditions, applied to the final master saliency maps.

For a more comprehensive comparison, we also input the AIM and SUN algorithms smaller (64×48), rescaled images, which greatly decreased their computational cost. These two variations are denoted as **AIM-small** and **SUN-small**. In Fig. 8, we show how the AUC score of all nine of these algorithms depends on the standard deviation of a Gaussian smoothing kernel applied to the final saliency maps.

From Fig. 8, we see that the performance of both RGB-Signature and Lab-Signature is very competitive with other saliency algorithms. The regions highlighted by the image signature saliency algorithm overlap to a surprisingly large extent with those image regions looked at by humans in free viewing. It is also interesting to observe that the optimal blurring factor σ is quite stable across different algorithms. In other words, we can choose one σ that works well for many algorithms. In Table 2, we list the AUC score of each algorithm under its optimal σ, as well as the mean σ

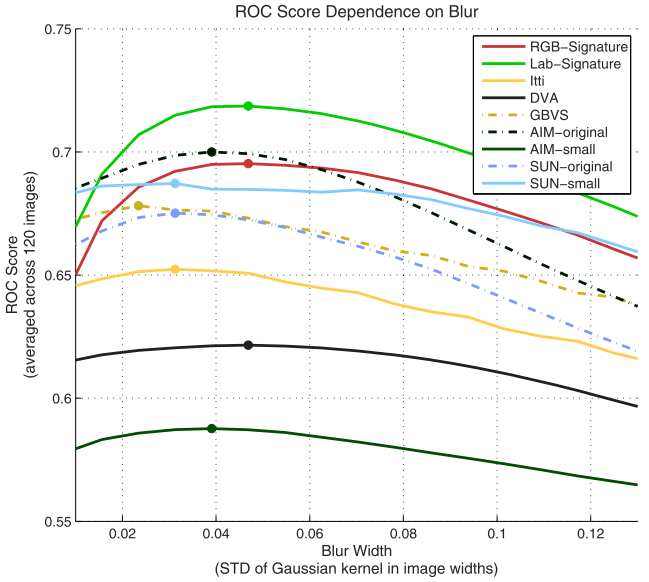

Fig. 8. The AUC metric reported herein and in other papers is quite sensitive to blurring. Parameterized by a Gaussian's standard deviation in image widths, this factor is explicitly analyzed to provide a better understanding of the comparative performance of an algorithm. For each algorithm, its optimal blurring factor is labeled as a dot on the plot. A solid line is used for algorithms whose average computing time is less than 1 second. For the computationally more expensive algorithms (GBVS, AIM-original, SUN-original), a dashed line is used to draw their performance curve.

(3.91 percent of the image width), together with average compute time per image in a standard desktop computing environment.

Importantly, not only is Lab-Signature the most predictive of fixations, it also runs faster than all competitors (except RGB-signature) in our tests of computational performance. This is due to its small number of channels and calculations compared to other saliency algorithms (see [14] for a comparison of saliency algorithms by computational components). Fig. 9 reports each algorithm's Matlab runtime measurements averaged over the data set. Compared to the image signature, which uses only three color channels at a single spatial scale, Itti and GBVS rely on seven feature channels and multiple spatial scales; DVA uses 192 filters of 192 dimensions, AIM uses 25 filters of 1,323 dimensions, and SUN uses 362 filters of 363 dimensions. Although these algorithms can be accelerated with efficient C implementations, the computational complexity of the image signature is lower, as suggested by the Matlab runtimes.

3.3 Correlations to Change Blindness

Change blindness [18] is a striking phenomenon in which a subject fails to notice otherwise obvious changes in a pair of images—even when the viewing time extends over a minute or longer. In such an experiment, the original image and a modified

TABLE 2
The Performance of All Nine Algorithms

Algorithm name	AUC (opt. σ)	AUC (mean σ)	time (s)
RGB-Signature	0.6952	0.6949	**0.024**
Lab-Signature	**0.7186**	**0.7183**	0.037
Itti	0.6524	0.6517	0.382
DVA	0.6216	0.6213	0.102
GBVS	0.6782	0.6760	2.060
AIM-original	0.7000	0.7000	8.303
AIM-small	0.5876	0.5876	0.131
SUN-original	0.6751	0.6745	9.885
SUN-small	0.6872	0.6849	0.580

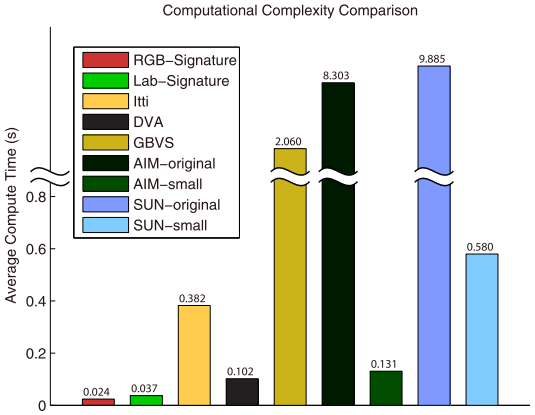

Fig. 9. Average compute time for each algorithm. All algorithms are implemented in the single-thread Matlab environment on an 8 core 2.5 GHz Xeon workstation.

version of it alternate repeatedly, but, critically, with a brief masking inserted in between. The ordinary perceptual motion or flicker which would accompany such a change is eliminated by the intervening interval which acts as a sort of mask. The observer must thus rely on his visual memory to identify the change. This is surprisingly difficult.

The phenomenon has inspired the rich literature in visual scene and object perception. Rensink et al. [19] suggest that an observer has to encode the image into an internal scene representation, which is sparse and incomplete. This very narrow bottleneck of representation has been demonstrated to be tightly related to the deployment of visual attention. There have been studies [20] that suggest that attended objects are more likely to be encoded in the working memory than nonattended ones.

Below, we use behavioral data from human subjects as an alternative ground truth to test the efficacy of our image signature. Results demonstrate that the signature distance of two image signatures is strongly (inversely) correlated with the reaction time of human subjects in detecting the change. To our knowledge, there have been no previous attempts to correlate a computational representation of a visual scene with the reaction time in a change blindness experiment.

3.3.1 Experiment Setup

In an experiment conceived by one of the authors (C.K.) and Claudia Wilimizig,[1] 60 color images of real-world scenes from personal albums were selected. For each original image, two modified versions were made, each with one object removed and retouched manually using Adobe Photoshop. The artifacts caused by image processing were kept minimal (Fig. 10 illustrates the experimental paradigm; Fig. 11 gives several examples of the stimuli). During each trial, the original image was displayed for 480 ms, followed by 160 ms black masking, and then 480 ms for the modified image, and then 160 ms masking. The trial stops after 60 s, or when the subject responds by clicking on the image. If the selected location was far away from the true modification, or if the subject did not respond within 60 s, or if the response time was less than 640 ms (before the first onset of the second image), the trial was discarded. Nine naive subjects with normal vision participated in the experiment. Subjects correctly identified the change (or signaled no change) in 1,011 (93.6 percent) of the $9 \times 2 \times 60 = 1{,}080$ trials.

Because the reaction time distribution among subjects is highly nonlinear, we instead compute the log reaction time. The inter-subject correlation (correlating one subject's reaction time against

1. Images were prepared by Amy Chung-Yu Chou and data were collected by Tom Laudes.

Fig. 10. The experimental paradigm for change blindness. In image 2, the window on the adobe wall has been removed. The subject has to report detection by clicking on the changed area of either image.

the remaining eight subjects) of the reaction times improves from 0.3558 to 0.5305 when moving from a linear to a log reaction time, suggesting that the log reaction time correlation is a more meaningful metric than the linear reaction time correlation.

3.3.2 Correlate Algorithm Output with Reaction Time

As a consequence of a complex cognitive process, the reaction time of a subject in a change blindness experiment is influenced by many factors. We here correlate such reaction times with various measures derived from the original image and its modified version.

First, reaction times are compared with the saliency of the modified objects. For a good saliency algorithm, we expect the saliency value of an object to be inversely correlated with the reaction time since the more salient an object is, the more easily a subject can spot it and thus detect its removal. The saliency value of a removed object is computed by the mean (or sum) pixel intensity of the object region in the saliency map of the original image.

Second, reaction times are compared to the Hamming distance (4) between the image signature descriptor of the original image and that of the modified image. As described in Section 2.3, this distance is a sensitive one when images share a background, as they do in the case of a change blindness pair. The distance between the descriptors should be related to the extent of difference in their salient, or foreground, regions.

Third, the widely used GIST descriptor [2] is used to describe each image in a change blindness pair, and reaction times are compared to the GIST distance. Torralba et al. [21] showed that perceptually similar images are usually close together in GIST descriptor space. GIST uses eight orientations, four scales for each 4×4 grid of an RGB color channel, mapping an image to a $8 \times 4 \times 16 \times 3 = 1{,}536$-dimensional real-valued descriptor.

Last, we use the pixel-wise distances between the images in the change blindness pair and compare these with reaction times. We actually use two pixel-wise measures: the ℓ_0 and ℓ_2 distances between the original and modified image. The ℓ_0 distance is exactly equal to the modified area size.

Let $\mathbf{h}_i$ be the log reaction times of the ith subject (a vector with a component for each image in the data set), $\mathbf{v}$ be the image pair distances according to one of the methods described above, then the normalized correlation c is given by correlating $\mathbf{v}$ with each $-\mathbf{h}_i$, normalized by the mean intersubject correlation, and averaging over nine subjects

$$c = \frac{1}{9}\sum_{i=1}^{9} \frac{\mathrm{corr}(-\mathbf{h}_i, \mathbf{v})}{E_{j \neq i}\left[\mathrm{corr}(\mathbf{h}_j, \mathbf{h}_i)\right]}. \tag{20}$$

The results are summarized in Fig. 12. Among all 13 algorithms, the Hamming distance between Lab-signature descriptors correlates best with reaction times. That is, among the methods tried here, the perceptual distance between change blindness pairs is best explained by the image signature descriptor. Given our understanding of the connection between foreground information and the signature, a difficult change blindness trial is likely one in which the removed object is perceived as part of the background, for in such a trial, we expect a small signature distance.

Fig. 11. Two sample image pairs. Labels indicate the median reaction time of nine subjects. Top: The difference is a small white post in the center divider (absent left, present right). Bottom: The difference is the yellow sign on the van (present left, absent right).

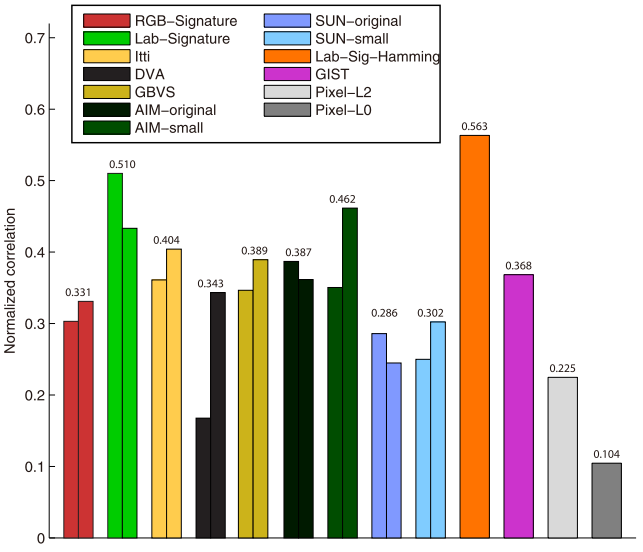

Fig. 12. The average normalized correlation between reaction time and algorithm outputs. For the first nine saliency algorithms, the left bar is the performance using the mean pixel value of the object region, whereas the right bar is the result of the sum of pixel saliency value (object size is variable). The score above each pair is the maximum correlation value among the two.

4 Conclusion

We introduced the image signature as a simple yet powerful descriptor of natural scenes. We proved on the basis of theoretical arguments that this descriptor can be used to approximate the spatial location of a sparse foreground hidden in a spectrally sparse background. We provided experimental data to show that the approximate foreground location highlighted by the image signature was remarkably consistent with the locations of human eye movement fixations, predicting them better than leading saliency algorithms at a fraction of the computational cost. We also provided results from a *change blindness* experiment in which the perceptual distance between slightly different images was predicted most accurately by the image signature descriptor.

Acknowledgments

The first author would like to thank Anders Hansen, Xiaodong Li, and Emmanuel Candes for their insightful discussions. We gratefully acknowledge Claudia Wilimzig, Amy Chung-Yu Chou, and Tom Laudes, who generated the images and collected data for the change blindness experiment. The research was supported by the NeoVision program at the US Defense Advanced Research Projects Agency (DARPA), by the US Office of Naval Research (via an award made through Johns Hopkins University), by the G. Harold & Leila Y. Mathers Charitable Foundation, and by the WCU (World Class University) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

References

[1] N. Bruce and J. Tsotsos, "Saliency Based on Information Maximization," *Proc. Advances in Neural Information Processing Systems,* pp. 155-162, 2006.
[2] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision,* vol. 42, no. 3, pp. 145-175, 2001.
[3] H. Zhou, H. Friedman, and R. von der Heydt, "Coding of Border Ownership in Monkey Visual Cortex," *J. Neuroscience,* vol. 20, no. 17, pp. 6594-6611, 2000.
[4] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *Arxiv preprint arXiv:0912.3599,* 2009.
[5] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[6] A. Oppenheim and J. Lim, "The Importance of Phase in Signals," *Proc. IEEE,* vol. 69, no. 5, pp. 529-541, May 1981.
[7] M. Hayes, J. Lim, and A. Oppenheim, "Signal Reconstruction from Phase or Magnitude," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 28, no. 6, pp. 672-680, 1980.
[8] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[9] E. Candes and T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?," *IEEE Trans. Information Theory,* vol. 52, no. 12, pp. 5406-5425, Dec. 2006.
[10] M. Rudelson and R. Vershynin, "On Sparse Reconstruction from Fourier and Gaussian Measurements," *Comm. Pure and Applied Math.,* vol. 61, no. 8, pp. 1025-1045, 2008.
[11] M. Goemans and D. Williamson, "Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming," *J. ACM,* vol. 42, no. 6, pp. 1115-1145, 1995.
[12] D. Ruderman, "The Statistics of Natural Images," *Network: Computation in Neural Systems,* vol. 5, no. 4, pp. 517-548, 1994.
[13] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual Correlates of Fixation Selection: Effects of Scale and Time," *Vision Research,* vol. 45, no. 5, pp. 643-659, 2005.
[14] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. Vision,* vol. 8, no. 7, pp. 1-20, 2008.
[15] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments," *Proc. Advances in Neural Information Processing Systems,* pp. 681-688, 2008.
[16] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Proc. Advances in Neural Information Processing Systems,* pp. 681-688, 2007.
[17] N. Bruce and J. Tsotsos, "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *J. Vision,* vol. 9, no. 3, pp. 1-24, 2009.
[18] D. Simons and M. Ambinder, "Change Blindness: Theory and Consequences," *Current Directions in Psychological Science,* vol. 14, no. 1, pp. 44-48, 2005.
[19] R. Rensink, J. O'Regan, and J. Clark, "To See or Not to See: The Need for Attention to Perceive Changes in Scenes," *Psychological Science,* vol. 8, no. 5, pp. 368-373, 1997.
[20] T. Kelley, M. Chun, and K. Chua, "Effects of Scene Inversion on Change Detection of Targets Matched for Visual Salience," *J. Vision,* vol. 3, no. 1, pp. 1-5, 2003.
[21] A. Torralba, R. Fergus, and Y. Weiss, "Small Codes and Large Image Databases for Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.