Object Separation in X-Ray Image Sets

Anonymous CVPR submission

Paper ID 983

Abstract

In the segmentation of natural images, most algorithms rely on the concept of occlusion. Intuitively, if all of the pixels in a region are the same color, they are probably part of the same object, which is blocking all of the light from the objects behind it. In x-ray images, however, this assumption is violated. In this paper, we introduce **SATIS**ϕ*, a method for separating objects in a set of x-ray images with the assumption of additivity in log space, where the log-attenuation at a pixel is the sum of the log attenuations of all objects that the corresponding x-ray passes through. Our method leverages multiple projection views of the same scene from slightly different angles to produce an accurate estimate of the extent and attenuation properties of objects in the scene. We demonstrate our algorithm on a set of collected x-ray scans, showing that our* **SATIS**ϕ *algorithm outperforms a standard image segmentation approach.*

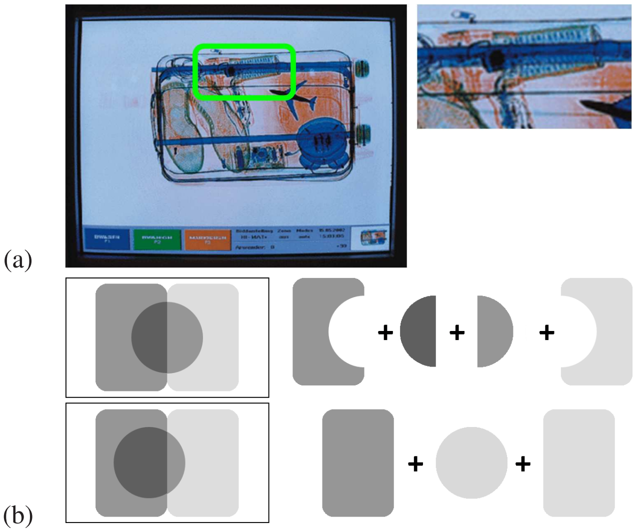


Figure 1. (a) An example baggage screening x-ray image. The object in the green box is partially obscured by the metal in the suitcase. The goal of this paper is to "see-through" the obscuring metal. (b) Synthetic example of the additivity of objects in transmission images.

1. Introduction

X-ray imaging is an important technology in many fields, from non-intrusive inspection of delicate objects, to weapons detection at security checkpoints [1]. Analysis of x-ray images in these applications shares many challenges with machine vision: we are interested in identifying "objects" and understanding their relations. For example, a security guard may search for an illegal substance in a suspected bag, or an archaeologist may inspect the content of an ancient artifact.

One particularly important application for x-ray scene analysis is in the field of *automatic threat detection*. Here, the aim is to build systems that can detect explosives concealed in bags using x-ray scans. Such systems have the potential to improve security checkpoints like the ones we meet at airports, but the general problem of scene understanding is clearly very hard. Fortunately, security screening outside the field of aviation, including offices, amusment parks and public transportation venues, involves bags that are far less cluttered than airplane carry-on bags, often containing only a handful of items. In these venues the main threats are massive bulks of explosives, hence detecion focuses on **estimating the chemical properties and the mass of each object, rather than its detailed shape**. As a result, developing a system that can separate a small number of overlapping objects has a huge potential to significantly improve automatic threat detection.

Most existing x-ray image analysis methods (*e.g.* [1]) use algorithms that were developed for visible spectrum images. These methods assume that objects are opaque and occlude other objects. While in the visible spectrum, light from objects behind the occluder are physically blocked from reaching the sensor, x-ray photons penetrate most materials. As a result, all objects along an x-ray path attenuate the x-ray, contributing to the final measured intensity. This is of course what allows x-ray imaging to "see-through" objects.

This transparency property has a fundamental effect on how x-ray images should be modeled and analyzed. Most importantly, unlike reflection images in which each pixel corresponds to a single object, *pixels in transmission images reflect the attenuation of multiple objects*. In the baggage x-ray of Figure 1(a), for example, the object in the green box is partially covered by the metal bar of of the suitcase. However, because the metal bar does not fully attenuate the x-rays, part of the attenuation in these pixels is due to the underlying item. In theory, we should be able to "subtract out" the metal bar, leaving a clear view of the object.

Our ultimate goal is to classify the objects in the bag. While many object classification works use parts-based models to recognize instances from a learned class [7], our objects of interest generally have neither the part structure, coherent shape, nor localized appearance that would be a good fit for these approaches. Instead, we consider a region based approach. Indeed, for many image analysis tasks, grouping pixels into regions for later classification is an effective techinque [4]. Standard approaches decompose an image into *disjoint* regions (segments) that should roughly correspond to objects. However, pixels in an x-ray image should not be assigned to a single region. For example, the pixels in the metal bar of of Figure 1(a) cannot be exclusively assigned to the suitcase or to the underlying object.

In this paper we develop a method to separate transmission images into potentially overlapping regions. The term *separation* distinguishes our output from a traditional segmentation, where each pixel belongs to a single region. The problem of decomposing a single image (sum of attenuations) into objects is ill-posed since it has more degrees of freedom than constraints. To address this problem we use information from multiple images to disambiguate the summed attenuation values (see Figure 1(b)). Computerized tomography (CT) takes this to the extreme, using thousands of scans to collect enough constraints to allow for a full 3D reconstruction. However, CT reconstruction formulation is highly sensitive to non-rigid or moving objects.

Here we take a different approach, and reduce the number of unknown variables. Our approach avoids the hard problem of fully reconstructing the image into a set of 3D objects, and focuses on identifying the rough shape and material of each object, performing well even with slightly deformable or moving objects. Our method is called **S**mall **A**ngle **T**ransmission **I**mage **S**eparation by ϕ (**SATIS**ϕ), and uses probabilistic priors and a reduced parameter space to make this problem tractable. **This paper focuses on the problem of identifying the composition of a scan in terms of the materials it contains**. To our knowledge this is the first attempt to address the extremely challenging problem of separating multiple overlapping and possibly deformable objects.

2. Related Work

In the x-ray community, the most common way of disambiguating objects is through CT reconstruction [13]. Such a reconstruction is typically obtained through the filtered back-projection algorithm [13] or algebraic reconstruction (ART) [3]. These approaches generally assume a large number of projection views are available, and that the scene being imaged is rigid during image acquisition. With a limited number of views, ART has been used somewhat successfully in previous work [3]. Unfortunately, ART breaks when objects can move between views, and is therefore not suitable for scans with liquids or moving parts.

In visible-spectrum images, structure from motion [5] and stereo vision [11] algorithms reconstruct the 3D scene using image data. These methods rely on occlusion, and will fail spectacularly for transmission images. A small number of these works [24, 12, 21] discuss 3D reconstruction in the presence of transparent objects. These works generally assume rigid objects and known camera geometry, and often rely on the geometry of light reflection or on active sensing methods, which are not applicable here.

Our goal of identifying "objects" is common in both the x-ray and visible-spectrum computer vision community. Indeed, our work is very closely tied to generic image segmentation. Typically, segmentation is the first step in a long processing chain [6]. Because each pixel has noise, grouping pixels allows for more robust processing. These segmentation algorithms have been used for object detection [8], scene categorization [20], content-based image retrieval [4], and other applications.

In the transparent object regime, [16] learns to recognize transparent objects based on texture, and [10] extracts the shape of transparent objects by projecting a light pattern onto the surface. Our work borrows many of the ideas from these application domains. [14] has used priors from natural images to separate an additive mixture of photos. Their approach is reported to be very sensitive to the presence of textures in the images (Fig5 therein). Finally, [2] and [23] both consider video sequences with transparent objects including reflections. Like us, they model the observed image as a sum of "layers." Both approaches use video sequences to resolve the layer ambiguity with an assumption of affine transformations between frames for each layer. In our case, however, we have much less data (only a few views), and the motion involves out-of-plane rotations, which may not be well-modeled by affine transformations.

3. Dual-Energy Projection X-ray Imaging

This section provides a short background on x-ray sensing, with a focus on dual-energy x-ray. This is the leading technique in security applications like explosive and drug detection. **Our objective is to identify the material composition of the obejcts that are present in an x-ray image**, or, more formally, the set of effective atomic numbers Z_o, and masses M_o for each object o. Figure 2(a) illustrates the setup for acquiring this data, including two views at small angle offsets from each other.

Projection x-ray imaging operates on the principle of intensity attenuation. An x-ray source emits a beam of x-ray photons with intensity I_0. As the photons pass through an object, they have a fixed probability per unit of length to interact with the material[1]. As a result, the intensity of the beam decays exponentially with a coefficient $\alpha(Z, E)$ that depends on Z, the atomic number of the object, and E, the energy of the x-ray photons [15]. The intensity detected at the sensor is therefore

$$I(E) = I_0 e^{-\alpha(Z,E)\rho t}, \tag{1}$$

where ρ is the density of the material, and t is the thickness of material that the ray passes through (in units of length). [2]. The values of the physical constants $\alpha(Z, E)$ were measured empirically for all relevant atomic numbers Z and energies E and are easily available [15]. However, our goal here is to extract the value of Z from a set of measured I (and the known parameters E and I_0 which are determined by the x-ray machine).

The value of Z cannot be isolated from a single measure of $I(E)$, since the exponent in Eq. (1) is a product of the α, ρ and t terms. To address this, dual-energy detectors are designed to measure separately the attenuation at two different energies E_1 and E_0. This allows us to cancel out the effect of ρt by considering the *dual-energy ratio* of logs

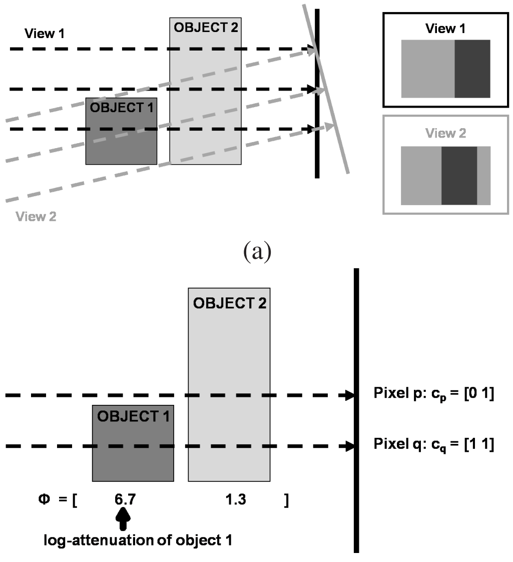
$$R = \frac{\log I(E_0)/I_0}{\log I(E_1)/I_0} = \frac{\alpha(Z, E_0)}{\alpha(Z, E_1)}, \tag{2}$$

where ρ and t cancel out because they do not depend on the x-ray energy. Using Eq. (2), we can solve for Z given the measured dual energy ratio R, and then backsolve for the product ρt. In the case of n objects, each object contributes multiplicatively to the final attenuation. The resulting log-attenuation is the sum of log-attenuations across objects

$$I = I_0 \prod_i e^{-\alpha(Z_i,E)\rho_i t_i} , \tag{3}$$

$$\log I(E)/I_0 = \sum_i^n \alpha(Z_o, E)\rho_o t_o \tag{4}$$

This *additivity* of the log-attenuations of individual objects allows us to develop efficient optimization algorithms for finding ϕ_i and features prominently in our **SATISϕ** model.



(a)

(b)

Figure 2. (a) Schematic representation of the acquisition of two views of a Small Angle Transmission Image (SATI) set. (b) Illustration of the composition variables c, indicating which objects a particular ray passes through, and the log-attenuation parameters ϕ, indicating the log-attenuation of each object in the scene.

4. The SATISϕ Model

Let $O = \{o_1, \ldots, o_n\}$ be a set of objects that is scanned at views $v_1, .., v_V$. The log attenuation value at a pixel p of view v is the sum of log-attenuation values over all objects that overlap p

$$\begin{aligned} \ell_{v,p} &= \sum_{o=1}^{n} c_{v,p,o}\phi_{v,p,o} + \xi, \\ c_{v,p,o} &= \begin{cases} 1 & \text{if } o \text{ overlaps } p \text{ in } v; \\ 0 & \text{otherwise.} \end{cases} \\ \phi_{p,v,o} &= \alpha(Z_o, E)\rho_o t_o \end{aligned} \tag{5}$$

where $c_{v,p,o}$ are the *composition variables* operating as “indicator” variable that select those objects that overlap with pixel p in view v. $\phi_{v,p,o}$ measures the log-attenuation at the pixel p of view v that is attributed to the object o (see Figure 2(b)). Added to this sum is ξ, a normally distributed noise vector $\xi \sim \mathcal{N}(0, \sigma_\ell^2 \mathbf{I})$ that reflects the imprecision in our model and in the measured data.

It is well known in the computer vision literature that grouping pixels into small regions (*superpixels*) yields more robust processing. We therefore applied a preprocessing step that aggregated pixels using the graph-based segmentation algorithm of [6]. In all the discussion below, we treat $\mathbf{c}$ as a vector over superpixels.

The problem of inferring the individual attenuations $\{\phi_{v,p,o}\}$ from a single scan measurement ℓ is clearly ill-posed, since there are more degrees of freedom than constraints. The standard solution is to collect thousands of scans of the objects, thus providing more constraints than degrees of freedom. In many cases however, the human visual system can separate a transmission image into ob-

[1] Assuming a single homogeneous object, and ignoring higher order effects (beam hardening)

[2] For non-homogeneous materials, the atomic number Z is replaced with the *effective atomic number*, Z_{eff}.

jects even with a small number of views. Figure 1(b) shows a motivating example. On the left are two images of the same set of objects. Assuming the visible-spectrum image model, the natural segmentation of the top image is shown by the top figure on the right. There are four regions, two rounded squares and two semicircles. However, if we assume that these are transmission images, the scene is most naturally represented by a full circle moving across two rounded squares. This decomposition is shown in the bottom right figure. Our goal is to develop a method for extracting this more intuitive decomposition using a small number of views, and a small number of real life objects.

Formulation as a Markov Random Field

We now reformulate the problem of object separation as a problem of probabilistic inference. We begin by reformulating Eq. (5) as a distribution

$$\ell_{v,p} \sim \mathcal{N}(\langle \mathbf{c}_{v,p}, \phi_{v,p} \rangle, \sigma_\ell^2), \tag{6}$$

where $\mathbf{c}_{v,p} \equiv [\mathbf{c}_{v,p,1}, \ldots, \mathbf{c}_{v,p,n}]$ and $\phi_{v,p} \equiv [\phi_{v,p,1}, \ldots, \phi_{v,p,n}]$ are vectors containing the indicator and log-attenuation values for all objects. Clearly, there are many values of $\phi_{v,p}$ and $\mathbf{c}_{v,p}$ that could together yield a maximal $\ell_{v,p}$, and additional constraints are needed to find a solution that would corresponds well to real objects.

We therefore introduce priors that favor decompositions that are more likely to occur. These priors, together with the probabilities for the observed attenuations variables ℓ induce a *Markov random field* (MRF) over the composition variables $\mathbf{c}$, with the log-attenuation values ϕ as real-valued "parameters." The data terms become the potentials

$$\psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi_{v,p}) = \mathcal{N}(\ell_{v,p}; \mathbf{c}_{v,p}^T \phi_{v,p}, \sigma_\ell^2). \tag{7}$$

We use priors in two forms: parameter equality constraints (parameter sharing), and MRF potentials over the composition variables. Our priors capture three properties of real scanned objects:

1. Object parts are homogeneous. We assume that objects are made of parts that have homogeneous material composition. A pair of scissors, for example, has a plastic handle and metal blades. In our model, we will treat each of these parts as separate "objects." Formally, this assumption implies that $\phi_{v,p,o} = \phi_{v,o}$ for all pixels that overlap the object o, allowing us to share these parameters.

2. Objects are compact. We assume that objects are continuous in space and as a result, if a pixel p overlaps an object, it is likely that all its neighbor pixels also overlap the object. This imposes a soft smoothness constraint on our objects, introducing the *smoothness potential*:

$$\psi_S(c_{v,p,o}, c_{v,q,o}) = \begin{cases} 1 & \text{if } c_{v,p,o} = c_{v,q,o}, \\ \gamma & \text{if } c_{v,p,o} \neq c_{v,q,o} \end{cases} \tag{8}$$

for all neighboring pixels p, q. In this potential, $\gamma < 1$ is a penalty suffered when the model has two neighboring pixels with different composition values.

3. Object attenuation changes smoothly across views. We assume that the given scans of the scene differ by only a small rotation angle $\theta \approx 0$, Figure 2(a). As a result, the effective thickness of each object varies as $cos(\theta) \approx 1$, yielding approximately equal attenuation for each view, $\phi_{v,o} = \phi_o, \forall v$. With this approximation, we have reduced the number of log-attenuation parameters down to the number of objects n.

Furthermore, since a small change in the scanning angle θ changes only slightly the silhouette of the object (and therefore the area in pixels), the area of an object should remain close to constant across views. We therefore introduce *area preservation potentials*:

$$\psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o}) = \exp\left(-(\mathbf{a}_v^T \mathbf{c}_{v,o} - \mathbf{a}_w^T \mathbf{c}_{w,o})^2 / 2\sigma_A^2\right), \tag{9}$$

where v and w are two views of the scene, and $\mathbf{a}_v$ is a vector containing the area (measured in raw pixels) of each superpixel in view v.

Combining these three types of potentials together, we obtain the **SATIS**ϕ MRF probability function:

$$\begin{aligned} Pr(\mathbf{c}; \phi) = & \frac{1}{Z} \prod_{v,p} \psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi_{v,p}) \qquad (10) \\ & \prod_{v,o,(p,q)} \psi_S(c_{v,p,o}, c_{v,q,o}) \prod_{v,w,o} \psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o}) \end{aligned}$$

where Z is a normalizing constant, known as the partition function.

This model has three image-independent parameters: σ_ℓ^2 – the noise variance in the image reconstruction potentials, γ – the smoothness penalty, and σ_A^2 – the variance of objects area across views. These parameters can be learned from data, but the scene-specific log-attenuation parameters ϕ must be estimated at test time for each image. This distribution trades off an assignment that faithfully represents the observed images, but also respects our smoothness and area preservation constraints.

To tune the scene-independent hyper parameters $\theta = \{\sigma_\ell^2, \gamma, \sigma_A^2\}$, we used a small "training set" of scenes to learn their values in a supervised way. First, a human annotator outlined the objects in the scene. Then we extracted the ground-truth composition variables $\mathbf{c}^{true}$, and the ML estimates of ϕ^{true} in each scene. Ideally, we would learn the maximum likelihood (ML) set of parameters θ given the assignment ϕ^{true}. However, since MRF learning is generally intractable [9], we use the simpler *piecewise training* scheme [22] where the parameters are estimated independently for each potential. This is equivalent to optimizing a lower bound to the partition function.

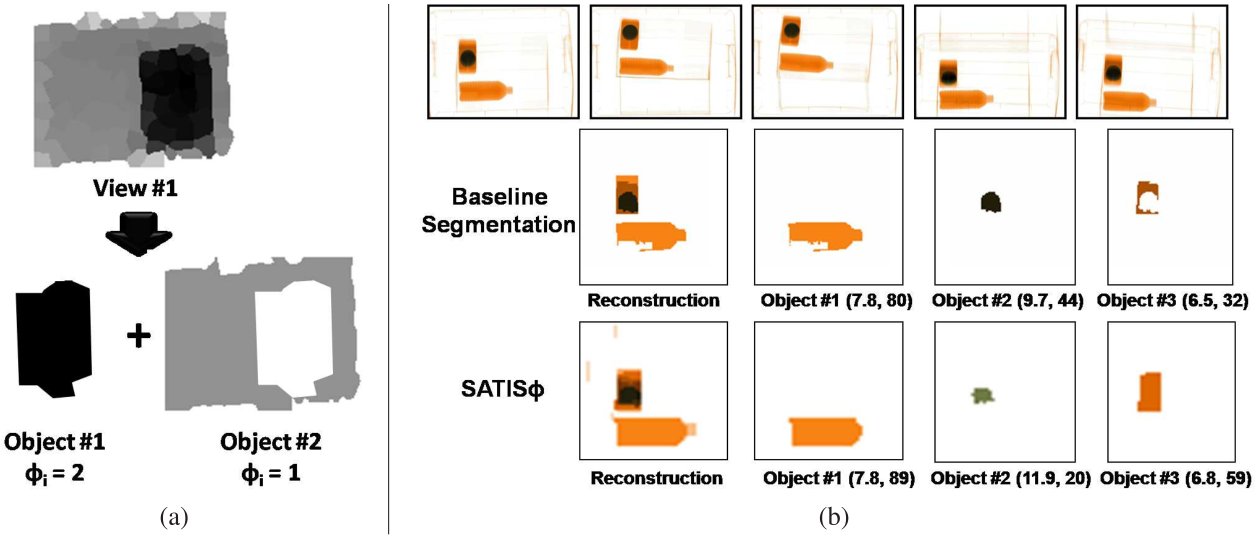


Figure 3. (a) An example showing where global moves, such as object splits, might be useful. (b) An example dataset used in this paper (top row), with the docomposition based on the segmentation of [6] (middle row), and the decomposition using our **SATIS**ϕ method. Objects are shown with the atomic number Z and mass M in parenthesis (Z,M). **SATIS**ϕ deduces both that Object #3 should have no hole, and that Object #2 is actually lighter because some of the attenuation is attributable to Object #3.

5. Optimization of the SATISϕ Decomposition

Given a dataset of SAT images for a scene, our goal is to find the *maximum a-posteriori* (MAP) decomposition according to our probabilistic model:

$$\begin{aligned}(\mathbf{c}^*, \phi^*) &= \text{argmax}_{\mathbf{c},\phi} Pr(\mathbf{c};\phi)\\ &= \text{argmin}_{\mathbf{c},\phi} -\log(Pr(\mathbf{c};\phi))\\ &= \text{argmin}_{\mathbf{c},\phi} \sum_{v,p} -\log\left[\psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi_{v,p})\right]\\ &\quad + \sum_{v,o,(p,q)} -\log\left[\psi_S(c_{v,p,o}, c_{v,q,o})\right]\\ &\quad + \sum_{v,w,o} -\log\left[\psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o})\right]. \qquad (11)\end{aligned}$$

We maximize this likelihood this with an algorithm in the spirit of Hard-EM. The algorithm alternates between finding a hard assignment to the hidden composition variables **c**, and finding the maximum likelihood estimate of ϕ given **c**. The objective is guaranteed to decrease at each step. However, we discovered that the likelihood manifold in this problem has a large number of local minima. We therefore prefer a view of the algorithm as a coordinate descent algorithm, and added a "global step" that modifies jointly both **c** and ϕ, to better escape local minima. We now describe these steps in more detail.

Initialization. We initialize the values of ϕ by finding homogeneous regions in a single image through a coarse segmentation. The parameters of this were tuned on a different dataset to produce segments that corresponded best to underlying objects.

Because the problem is non-convex, the initialization is likely to have a large effect on convergence to local minima. We therefore tested four initialization strategies, all fully-unsupervised: a) The segmentation of [6] worked best, b) The algorithm of Ren and Malik [19], c) setting ϕ based on observed attenuations, and d) a kmeans segmentation.

Iterations. After initialization, we iterate through three steps: (1) optimization of the composition variables **c** with a fixed ϕ, (2) optimization of the log-attenuation vector ϕ with a fixed **c**, and (3) move-based global optimization.

1. Optimization of c. Given ϕ, Eq. (11) can be rewritten as a quadratic problem in $c_{v,p,o}$.

$$\begin{aligned}\min_{\mathbf{c}} \quad & w_1\|\mathbf{P}\mathbf{c} - \ell\|^2 + w_2\|\mathbf{S}\mathbf{c}\|^2 + w_3\|\mathbf{D_A}\mathbf{c}\|^2\\ \text{s.t.} \quad & \mathbf{c}_{v,p,o} \in \{0,1\} \quad , \qquad (12)\end{aligned}$$

where **c** is a vector that contains all elements of $\mathbf{c}_{v,p,o}$, **P** is a matrix with rows $\mathbf{p}_i^T$ such that $\mathbf{p}_i^T\mathbf{c} = \mathbf{c}_{v,p}^T\phi = \ell_{v,p} + \xi$ (this term corresponds to ψ_ℓ), **S** is a matrix that computes the difference in c_p and c_q for each neighbor pair (this term corresponds to ψ_S), and $\mathbf{D_A}$ is a matrix that computes the objects' area differences across views (this term corresponds to ψ_A). The weights (w_1, w_2, w_3) are computed from the scene-independent parameters.

This problem is an integer program with a convex (quadratic) objective. We find an approximate solution to this problem with an interative relax-and-round procedure. In the relaxation phase, we use a *convex relaxation* approach that was shown to be extremely effective in MAP inference in MRFs [18]. We relax the integer constraints by replacing the binary variables **c** with $\tilde{\mathbf{c}} \in I\!R$, and the constraints in Eq. (12) with: $0 \le \tilde{c}_i \le 1$. The resulting problem is a

Quadratic Program (QP) that can be solved efficiently.

In the rounding phase, we look at the real solution $\tilde{c}$ and select the largest values for each "object." For each such value that is above some threshold R (we use $R = 0.5$ in experiments below), we set it's value to 1, and "freeze" the optimization variable. We then iterate, re-solving a new QP with a subset of the values frozen to 1. At each iteration, more composition variables (c's) are turned on. When no more variable can be set to 1, we round the remaining values using the procedure of [18].

2. Optimization of ϕ. Given $\mathbf{c}$, the objective Eq. (11) depends on ϕ only through a squared-error term for each ℓ. Minimizing the cost with respect to ϕ is a linear least-squares problem.

3. Joint Optimization of $(\mathbf{c}, \phi)$. To reduce the problem of local minima, we added four types of "global moves" that change both $\mathbf{c}$ and ϕ simultaneously: merges, removals, object splits and component splits. Figure 3(a), illustrates object splits using a local minimum case where no isolated change to $\mathbf{c}$ or ϕ improves the objective. Splitting object #1 (with $\phi_1 = 2$) so that its pixels belong both to object #2 (with $\phi_2 = 1$) and a *new* object, with $\phi_3 = 1$, improves the objective by yielding a smoother object #2.

Greedy completion. Once iterating has converged, we finish with a greedy descent stage which optimizes the composition variables $\mathbf{c}_{v,p}$ in sequence. We sweep through each superpixel p in each view v, and enumerate all the possible values for the vector $\mathbf{c}_{v,p}$ (2^n assignments for n objects), plus the corresponding optimal ϕ vector, and select the setting with the lowest cost. This step serves to fix errors introduced by rounding the composition variables.

6. Experimental Results

We tested the **SATIS**ϕ method on a collection of 23 datasets of SAT image sets, each collected with a dual-energy x-ray *Astrophysics* machine. We selected 3 of these datasets to serve as training data, and used them to estimated the scene-independent parameters: $\sigma_\ell^2 = 0.29, \gamma = 0.32, \sigma_A^2 = 100.2$.

For comparison with a **baseline** method, we segmented the image using the method of [6], with no further processing. This is a state-of-the-art approach for extracting object-parts in a natural image. Figure 3(b, 2^{nd} row) shows an example **baseline** decomposition. To the best of our knowledge, there is no other competing strategy in the literature.

We ran the **SATIS**ϕ decomposition algorithm on each of the other 20 datasets. Figure 5(a) shows a decomposition obtained across all views of a sample test instance, and Figure 5(b) and (c) show the decomposition for a single view. In all three instances, the objects appear to be large and to generally correspond to the objects present in the scene.

To understand the difference between the operation of **SATIS**ϕ and the baseline segmentation, compare the decompositions in Figure 3(b). **SATIS**ϕ correctly identifies that the vertical orange object should form a continuous rectangle, whereas the **baseline** decomposition has a hole in the object. Since **baseline** attributes the entire attenuation of the circular segment to object #2, the atomic number is estimated as 9.7 (nearly organic). **SATIS**ϕ correctly identifies that the attenuation in these pixels is due to both objects #2 and #3, and therefore the circular object has atomic number of 11.7, corresponding to light metals.

While obtaining the correct visual decomposition is important in itself, the more important goal of this analysis is to determine the material properties of the objects present. As described in Section 3, for each object, we can convert from log-attenuation values into atomic number Z and mass approximation M. These properties will be used for further analysis. As a post-processing step to make this comparison more informative, we remove extracted objects with mass smaller than 2.5 per view. Because the baseline segementation often produces tiny segments, this allows us to only focus on objects of reasonable size.

For each scan in the test set, we used hand-annotation of the objects to compute the "ideal" material mass and atomic number that can be obtained from this data. Each object produces a single point in *MZ space* (mass – atomic number space) where the y-coordinate corresponds to the atomic number, and the x-coordinate corresponds to the mass attenuation of the object. The quality of a decomposition can be quantitatively evaluated by measuring how "close" the extracted objects are to the true objects in MZ space.

Figure 5(d,e,f) plot the extracted objects from the two methods agains the groundtruth objects. These plots correspond to the decompositions of data sets (a),(b), and (c), and plot the hand-annotated objects, the objects obtained by the **SATIS**ϕ method and the segmentation baseline objects in MZ space. In the example of Figure 5(a), the two metal objects (2 and 3) are nearly perfectly extracted, while the organic (orange) rectangle is split between three objects (1, 4, and 5). The graph of (d) indicates this, by showing the two blue stars that match the red stars in the high Z region and the three organic blue stars that together add up to the mass of the single organic red star in the low Z region. In general, the **SATIS**ϕ objects tend to lie closer to the groundtruth objects than the **baseline** ones do.

In Table 4 we quantify these results for the entire dataset. For each hand-annotated object, we match it to the closest extracted object. If the Euclidean distance (in MZ space) is less than T_{miss}, we accept the match. Column 3 provides the number of matches for both methods for a range of thresholds. Once we have obtained a match, we are also interested in how accurate it is. To investigate this, we find all the "common" matches (true objects matched by **SATIS**ϕ and the **baseline**), and show the number of such matches in column 4. Columns 5, 6, and 7 of Table 4

Method	T	Matches	Common	RMS mass-Z	RMS mass	RMS Z	Misses	Extra
SATISϕ	0.10	29	16	0.0541	0.0420	0.0340	37	36
Baseline	0.10	26	16	0.0701	0.0561	0.0420	40	64
SATISϕ	0.05	14	1	0.0025	0.0015	0.0020	52	51
Baseline	0.05	10	1	0.0500	0.0500	0.0020	56	80
SATISϕ	0.02	6	0	N/A	N/A	N/A	60	59
Baseline	0.02	2	0	N/A	N/A	N/A	64	88

(a)

Noise	Matches
0	29

(b)

Figure 4. Quantitative evaluation of the match between extract objects and hand-annotated object. We show the root-mean-squared (RMS) distances for matched objects in MZ space (mass – atomic number). We also show the number of matches, misses, and extra objects extracted, for three different settings of matching threshold T_{miss}.

show the root-mean-squared (RMS) distance in MZ space, mass alone, and Z alone between the matching extracted and hand-annotated objects for the common matches. For each setting, the **SATISϕ** objects have better RMS distances to the hand-annotated objects than the baseline.

Annotated objects that produce no match are "misses," and unmatched extracted objects are "extras." Our baseline tends to produce more objects than **SATISϕ**, leading to both more misses (col. 8), and more extras (col. 9). This analysis shows that even with more "guesses" for objects, the baseline still cannot do as well as **SATISϕ**.

In our final experiment, we explored the robustness of our method to imaging noise. Such noise may be produced by either errors in the data acquisition setup, or by background clutter such as clothing, which is less dense and more heterogeneous than the objects we consider in our data. For each of our 20 test datasets, we add white noise with standard deviation σ_N^2 to both the log-high and log-low image channels. To evaluate the stability, we look at the number of matches at each noise level. Figure **??** shows the results. *****GAH: Write something smart about these results here*****

The primary error mode appears to be oversplitting of true objects across layers. This is apparent in the example of Figure 5(a), where the rectangular organic object is split into extracted objects 1, 4, and 5, and in Figure 5(b), where the electronic box is incorrectly split between objects 1 and 2. Despite this, we still achieve much larger and more realistic object approximations than the baseline.

7. Discussion

In this paper, we have shown that transmission images can be decomposed into the objects that make up the image "layers.". The **SATISϕ** model, successfully disambiguates the objects with a small number of views of the same scene. We have described an efficient search method for finding a high-scoring decomposition, and have showed the method's effectiveness on real x-ray scans.

Standard x-ray processing approaches (CT) aim to provide exact 3D reconstruction, and are limited to rigid and stationary objects. The probabilistic approach taken here allows us to extract the crucial information: the chemical composition of the objects. It uses only a handful of scans and is robust against objects that are slightly deformable. In fact, **SATISϕ** actually benefits from objects that move relative to each other, in the same way that humans do when asked to interpret a scene with transparent objects [17].

In addition, the probabilistic approach taken here allows to introduce priors that penalize solutions that are physically unrealistic. We tested simple smoothness and area preservation priors, but more complex priors may be introduced to improve the accuracy of the decomposition. Specifically, as in the work of [2], we can identify object junctions (edges or corners) to provide additional information about object correspondences across views. In addition, the **SATISϕ** smoothness potential treats each pair of neighbors equally. Using some information from the image, such as the presence or absence of a strong image edge, is likely to improve the precision.

The **SATISϕ** model was designed to handle separating objects in x-ray images, but the underlying ideas could apply to more general problems. For instance, similar techniques could be taken for segmenting semi-transparent objects [16] and reflections [14].

References

[1] B. Abidi, J. Liang, M. Mitckes, and M. Abidi. Improving the detection of low-density weapons in x-ray luggage scans using image enhancement and novel scene-decluttering techniques. *Journal of Electronic Imaging*, 13(3):523–538, 2004. 1

[2] E. Adelson. Layered representations for vision and video. In *Proc. of the IEEE WS on Representation of Visual Scenes, Cambirdge, MA*, page 3, 1995. 2, 7

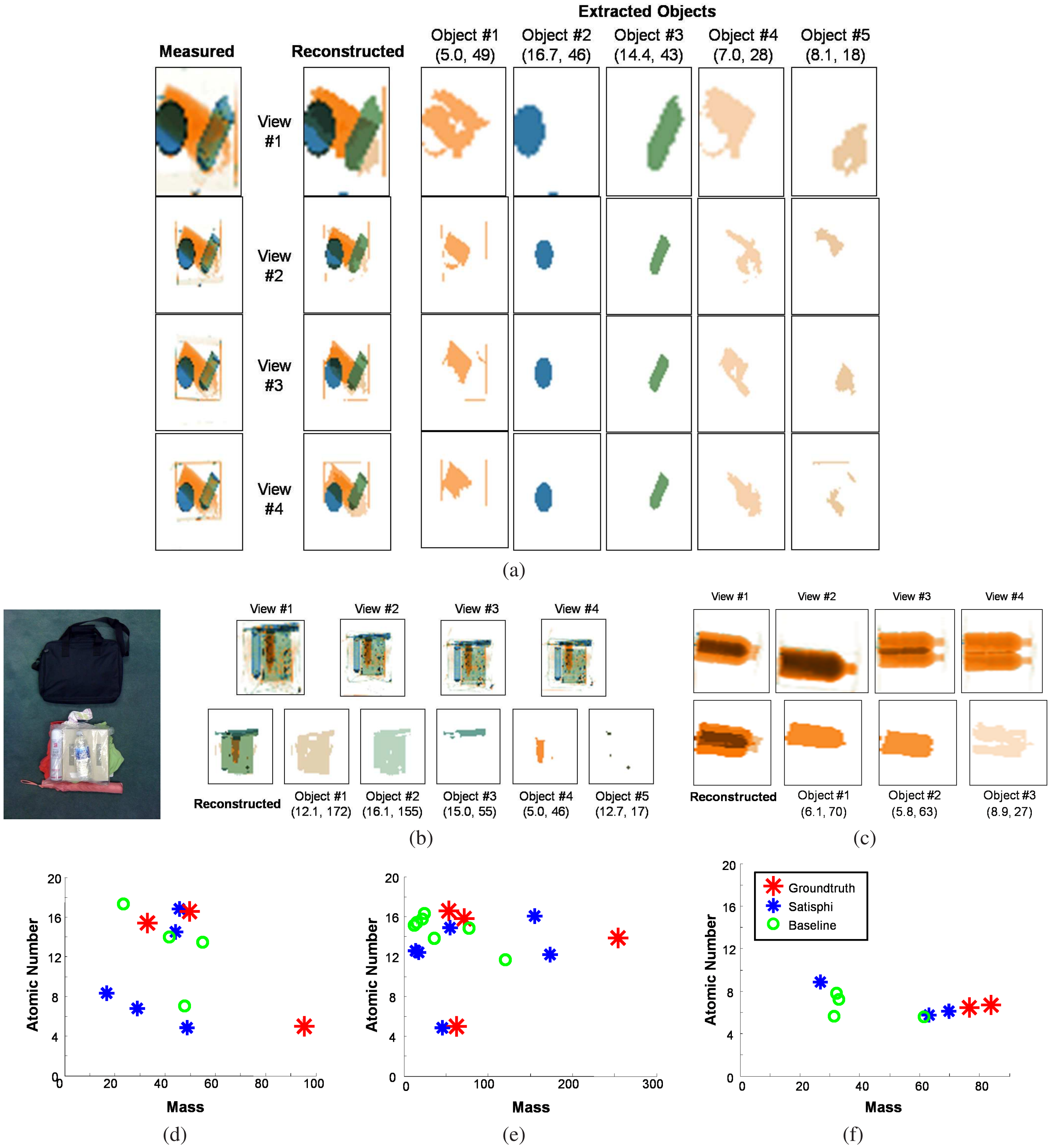
[3] A. H. Andersen. Algebraic reconstruction in ct from limited views. *IEEE Transactions on Medical Imaging*, 8(1):50–55, 1989. 2

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1026–1038, 1999. 2

[5] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from motion without correspondence. In *Computer Vision and Pattern Recognition*, June 2000. 2

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 2, 3, 5, 6

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsuper-



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. (a) Reconstructions obtained by optimization of the **SATIS**ϕ objective. The left column shows the original x-ray scans of a scene. The second column shows the **SATIS**ϕ reconstruction , and the remaining columns correspond to individual objects in the **SATIS**ϕ decomposition. Objects are colored to emphasize the material of the scanned objects, and are labeled with the atomic number Z and mass M in parenthesis (Z,M). (b,c) Reconstructions for two other scenes in the test set, showing the original images (top row), and the reconstruction and decomposed objects for a single view (bottom row). To the left of (b), we show a visible spectrum photograph of the contents of the bag. (d,e,f) Plots of the the objects discovered in the decompositions of (a),(b), and (c) in MZ (mass – atomic number) space.

vised scale-invariant learning. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:264, 2003.

[8] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. *Computer Vision and Pattern Recognition*, 0:1–8, 2008. 2

[9] V. Ganapathi, D. Vickrey, J. Duchi, and D. Koller. Constrained approximate maximum entropy learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008. 4

[10] S. Hata, Y. Saitoh, S. Kumamura, and K. Kaida. Shape extraction of transparent object using genetic algorithm. In *International Conference on Pattern Recognition*, volume 4, page 684, 1996. 2

[11] R. D. Henkel. A simple and fast neural network approach to stereovision. In *Neural Information Processing Systems*, pages 808–814, 1998. 2

[12] I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich. State of the art in transparent and specular object reconstruction. In *STAR Proceedings of Eurographics*, 2008. 2

[13] A. C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. 2

[14] A. Levin, A. Zomet, and Y. Weiss. Learning to perceive transparency from the statistics of natural scenes, 2002. 2, 7

[15] D. Lide. *CRC handbook of chemistry and physics*. CRC press, 2004. 3

[16] K. McHenry, J. Ponce, and D. Forsyth. Finding glass. In *Computer Vision and Pattern Recognition*, pages 973–979, 2005. 2, 7

[17] F. Metelli. The perception of transparency. *Scientific American*, 230:91–98, 1974. 7

[18] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *International Conference on Machine learning*, pages 737–744, 2006. 5, 6

[19] X. Ren and J. Malik. Learning a classification model for segmentation. In *Proc. 9th Int'l. Conf. Computer Vision*, volume 1, pages 10–17, 2003. 5

[20] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[21] S. Soatto and P. Perona. Three dimensional transparent structure segmentation and multiple 3d motion estimation from monocular perspective image sequences. In *International Conference on Pattern Recognition*, 1994. 2

[22] C. A. Sutton and A. McCallum. Piecewise training for undirected models. In *Uncertainty in Artificial Intelligence*, pages 568–575, 2005. 4

[23] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:1246, 2000. 2

[24] M. Yamazaki, S. Iwata, and G. Xu. Dense 3d reconstruction of specular and transparent objects using stereo cameras and phase-shift method. In *Asian Conference on Computer Vision (ACCV)*, 2007. 2