Object Separation in X-Ray Image Sets

Geremy Heitz and Gal Chechik*
Qylur Security Systems, Inc.
Palo Alto, CA 94305
`gheitz,gchechik@qylur.com`

Abstract

In the segmentation of natural images, most algorithms rely on the concept of occlusion. In x-ray images, however, this assumption is violated, since x-ray photons penetrate most materials. In this paper, we introduce SATISϕ, a method for separating objects in a set of x-ray images using the property 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 attenuation properties of objects in the scene. These properties can be used to identify the material composition of these objects, and are therefore crucial for applications like automatic threat detection. We evaluate SATISϕ on a set of collected x-ray scans, showing that it outperforms a standard image segmentation approach and reduces the error of material estimation.

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 *automatic threat detection*. Here, the aim is to 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. In this paper we focus on **estimating the chemical properties and the mass of each object, rather than its detailed shape**. For reasons detailed below, handling overlapping objects is perhaps the major barrier to accurate detection. Developing a system that can separate overlapping objects has the potential to significantly improve automatic threat detection. However, there is no published work that we are aware of on separating overlapping objects in x-ray scans.

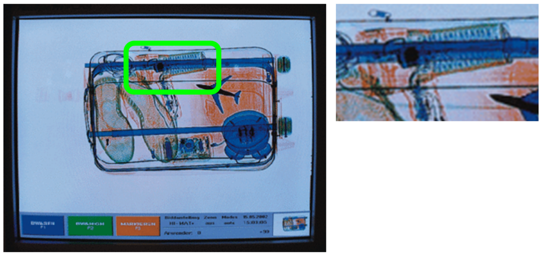


Figure 1. Example x-ray image of baggage. The object in the green box is partially obscured by the metal in the suitcase. The goal of this paper is to "see-through" the obscuring metal.

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 each other. X-ray photons, however, penetrate most materials. As a result, all objects along an x-ray path attenuate the x-ray and contribute to the final measured intensity. This is 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 represent 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. For regular images, machine vision approaches decompose an image into *disjoint* regions (segments) that should roughly correspond to objects. For x-ray images, however, pixels should not be assigned to a single region, but rather to all overlapping objects that contribute to their attenuation.

---

[*] Also at Google Research, Mountain View, 94043, CA, and Gonda Brain Research center, Bar Ilan University, Ramat Gan, 52900 Israel

In this paper we develop a method to separate transmission images into potentially overlapping regions. The term *separation* distinguishes our output from traditional segmentation, where each pixel belongs to a single region. The problem of decomposing a single image into objects is ill-posed, with more degrees of freedom than constraints. To address this problem we use information from multiple images to disambiguate the summed attenuation values. 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 is highly sensitive to non-rigid or moving objects.

Here we take a different approach and reduce the number of unknown variables instead of increasing the number of constraints. Our approach, named SATISϕ, 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. It uses probabilistic priors and a reduced parameter space to make this problem tractable. To our knowledge this is the first attempt to address the extremely challenging problem of separating multiple overlapping and possibly deformable objects in x-ray data.

2. Related Work

In the x-ray community, a common way of disambiguating objects is through CT reconstruction [12]. This is typically obtained through the filtered back-projection algorithm [12] 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. Taking a different approach, [15] overcomes the lack of enough measurement data by applying compressed sensing principles in the form of sparsity constraints to their MRI reconstructions. Similarly, we rely on the use of prior constraints to combat lack of data.

In visible-spectrum images, structure from motion [25] and stereo vision [10] 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 [26, 11, 22] 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 [5]. Because each pixel has noise, grouping pixels allows for more robust processing. These segmentation algorithms have been used for object detection [7], scene categorization [21], content-based image retrieval [4], and other applications.

Many works on object classification in regular images use parts-based models to recognize instances from a learned class [6]. 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 technique [4].

In the transparent object regime, [16] learns to recognize transparent objects based on texture, and [9] 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. [13] 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 (Fig 5 therein), and is not expected to work well in the current setting. Finally, [2] and [24] 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 objects 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 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[2], and E, the energy of the x-ray photons [14]. The intensity detected at

---

[1] For a homogeneous material, ignoring beam hardening effects.

[2] For non-homogeneous materials, the atomic number Z is replaced with the *effective atomic number*, Z_{eff}, which is approximately the weighted average of the atomic numbers of the component materials.

the sensor is therefore

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

where ρ is the density of the material, I_0 is the initial intensity, and t is the thickness of material that the ray passes through (in units of length)[2]. Our goal 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 measurement of $I(E)$, since the exponent in (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_2. This allows us to cancel out the effect of ρt by considering the *dual-energy ratio* of logs

$$R = \frac{\log I(E_2)/I_0}{\log I(E_1)/I_0} = \frac{\alpha(Z,E_2)}{\alpha(Z,E_1)}. \tag{2}$$

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 [14]. This allows us to solve (2), 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. $I = I_0 \prod_{o=1}^{n} e^{-\alpha(Z_i,E)\rho_i t_i}$. The resulting log-attenuation is the sum of log-attenuations across objects

$$-\log I(E)/I_0 = \sum_{o=1}^{n} \alpha(Z_o,E)\rho_o t_o \tag{3}$$

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.

4. The SATISϕ Model

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

$$\begin{aligned} \ell_{v,p} &= \sum_{o=1}^{n} c_{v,p,o}\phi_{v,p,o} + \xi, \qquad (4) \\ c_{v,p,o} &= \begin{cases} 1 & \text{if } o \text{ overlaps } p \text{ in } v; \\ 0 & \text{otherwise.} \end{cases} \\ \phi_{p,v,o} &= \alpha(Z_o,E)\rho_o t_o \end{aligned}$$

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, and are illustrated in figure 2. $\phi_{v,p,o}$ measures the log-attenuation at the pixel p of view v that is attributed to the object o. 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.

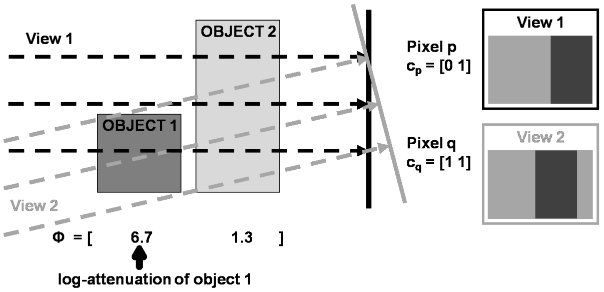


Figure 2. Acquisition schematic for two views of a scene. The composition variables c for two pixels are shown to the right, and the log-attenuation parameters ϕ for each object are shown below. The two resulting images are shown in the right column.

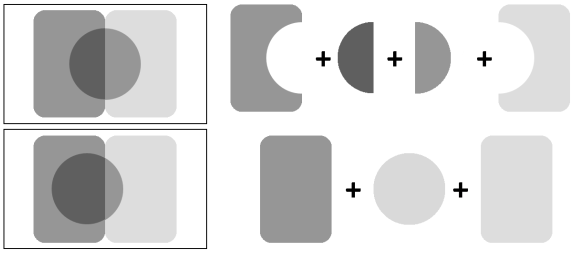


Figure 3. Illustration why using few view helps to disambiguate objects and select the correct separation in transmission images.

It is well known that grouping pixels into small regions (*superpixels*) often yields more robust processing. We therefore applied a preprocessing step that aggregated pixels based on their attenuation values and using the graph-based segmentation algorithm of [5]. In all 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 increase the number of constraints by collecting thousands of scans. In many cases however, the human visual system can separate a transmission image into objects even with a small number of views. Figure 3 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. Building on this intuition, we now describe a method aiming to decompose transmission images using a few views of a few 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 writing

(4) as a distribution

$$\ell_{v,p} \sim \mathcal{N}(\langle \mathbf{c}_{v,p}, \phi_{v,p} \rangle, \sigma_\ell^2), \tag{5}$$

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

$$\psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi_{v,p}) = \mathcal{N}(\ell_{v,p}; \mathbf{c}_{v,p}^T \phi_{v,p}, \sigma_\ell^2). \tag{6}$$

We use priors in two forms: parameter equality constraints, 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 p 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, its neighbor pixels (q's) are more likely to overlap the object as well. This imposes a soft smoothness constraint on our objects, introducing the *smoothness potential*:

$$\psi_S(c_{v,p,o}, c_{v,q,o}) = \begin{cases} 1 & \text{if } c_{v,p,o} = c_{v,q,o}, \\ \gamma & \text{otherwise} \end{cases} \tag{7}$$

for all neighboring pixels p, q. $\gamma < 1$ is a penalty suffered when neighboring pixels have different compositions.

3. Object attenuation changes smoothly across views. We assume that the scans of the scene differ by only a small rotation angle $\theta \approx 0$ (see 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, $\phi_{v,o} = \phi_o, \forall v$. This approximation reduces the number of log-attenuation parameters down to the number of objects n. It could also be relaxed into a soft penalty.

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*:

$$\psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o}) = \exp\left(-(\mathbf{a}_v^T \mathbf{c}_{v,o} - \mathbf{a}_w^T \mathbf{c}_{w,o})^2 / 2\sigma_A^2\right), \tag{8}$$

where v,w are two views, 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 SATISϕ MRF probability function:

$$\begin{aligned} Pr(\mathbf{c}; \phi) = & \frac{1}{Z} \prod_{v,p} \psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi) \\ & \prod_{v,o,(p,q)} \psi_S(c_{v,p,o}, c_{v,q,o}) \prod_{v,w,o} \psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o}) \end{aligned} \tag{9}$$

where Z is a normalizing constant (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 object 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 a decomposition that faithfully represents the observed images, but also respects our smoothness and area preservation constraints.

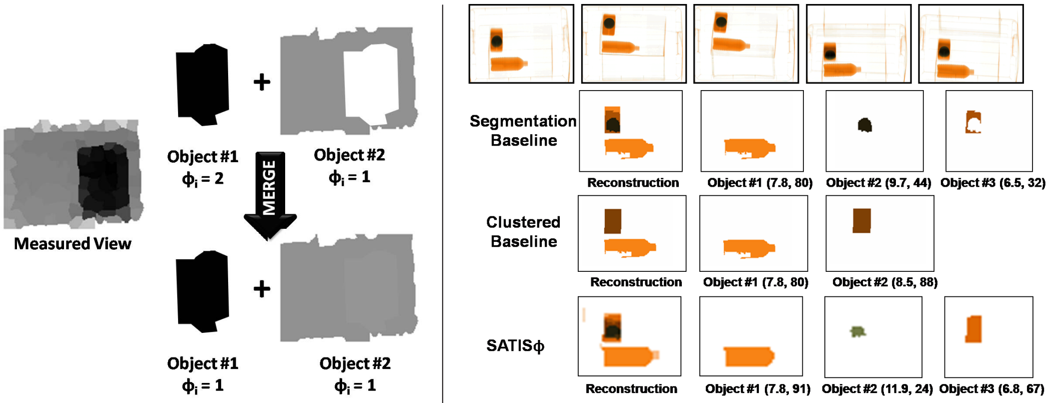
To tune the scene-independent hyper parameters $\theta = \{\sigma_\ell^2, \gamma, \sigma_A^2\}$, we used a small "training set" of scenes to learn their values in a supervised way. First, a human annotator outlined the objects in the scene. Then we extracted the ground-truth composition variables $\mathbf{c}^{true}$, 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 [8], we use the simpler *piecewise training* scheme [23] where the parameters are estimated independently for each potential. This is equivalent to optimizing a lower bound to the partition function.

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:

$$\begin{aligned} (\mathbf{c}^*, \phi^*) = & \ \text{argmax}_{\mathbf{c},\phi} Pr(\mathbf{c}; \phi) \\ = & \ \text{argmin}_{\mathbf{c},\phi} \sum_{v,p} -\log\left[\psi_\ell(\ell_{v,p}, \mathbf{c}_{v,p}; \phi_{v,p})\right] \\ & + \sum_{v,o,(p,q)} -\log\left[\psi_S(c_{v,p,o}, c_{v,q,o})\right] \\ & + \sum_{v,w,o} -\log\left[\psi_A(\mathbf{c}_{v,o}, \mathbf{c}_{w,o})\right]. \end{aligned} \tag{10}$$

We maximize this likelihood with an algorithm in the spirit of Hard-EM. The algorithm alternates between finding a hard assignment to the hidden composition variables $\mathbf{c}$, and finding the maximum likelihood estimate of ϕ given $\mathbf{c}$. While the objective is guaranteed to decrease at each step, we discovered that the likelihood manifold has a large number of local minima. We therefore prefer a view of the algorithm as a coordinate descent algorithm, and added "global



(a) (b)

Figure 4. (a) A global “merge” move. By simultaneously changing the log-attenuation of Object #1 and the spatial extent of Object #2, we achieve a high probability separation, because Object #2 becomes “smoother”. Here, the split immediately reduces the smoothness penalty, and allows for later steps to further improve the total score. (b) An example dataset used in this paper (top row), with the decomposition based on the segmentation of [5] (second row), based on the X-Means clustering of these segments (third row), and using our SATISϕ method (bottom row). Objects are shown with the atomic number Z and mass M in parenthesis (Z,M). SATISϕ 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.

steps” that modifies jointly both $\mathbf{c}$ and ϕ, to better escape local minima. We now describe these steps in detail.

Initialization. We initialize ϕ by finding homogeneous regions in a single image through a coarse segmentation. Since the problem is non-convex, the initialization is likely to significantly impact convergence to local minima. We therefore tested four initialization strategies, all unsupervised: a) The segmentation of [5], which worked best, b) The algorithm of Ren and Malik [20], c) setting ϕ based on observed attenuations, and d) a k-means segmentation.

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

1. Optimization of c. Given ϕ, (10) can be rewritten as a quadratic problem in $c_{v,p,o}$.

$$\begin{aligned} \min_{\mathbf{c}} \quad & w_1\|\mathbf{P}\mathbf{c}-\ell\|^2 + w_2\|\mathbf{S}\mathbf{c}\|^2 + w_3\|\mathbf{D_A}\mathbf{c}\|^2 \\ \text{s.t.} \quad & \mathbf{c}_{v,p,o} \in \{0,1\} \quad , \end{aligned} \tag{11}$$

where $\mathbf{c}$ is a vector that contains all elements of $\mathbf{c}_{v,p,o}$, $\mathbf{P}$ 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), $\mathbf{S}$ 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 iterative 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 [19]. We relax the integer constraints by replacing the binary variables $\mathbf{c}$ with $\tilde{\mathbf{c}} \in I\!R$, and the constraints in (11) 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 its 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 [19].

2. Optimization of ϕ. Given $\mathbf{c}$, the objective (10) 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 4(a) illustrates object merges using a local minimum case where no isolated change to $\mathbf{c}$ or ϕ improves the objective. Merging Object #1 ($\phi_1 = 2$) with Object #2 so that its pixels belong both to #1 and #2 (with $\phi_2 = 1$ and the *new* $\phi_1 = 1$) 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 SATISϕ method on a collection of 23 datasets of SAT images collected with a dual-energy x-ray *Astrophysics* machine. These datasets were constructed to mimic the distribution of bags in public venues like amusement parks or public transit stations. Bags in these venues are typically less cluttered than airplane carry-on bags, often containing only a handful of items. Also, in these venues the main threats are massive bulks of explosives, where automatic detection is likely to be more successful.

Each of the 23 datasets contained 5 scans of 2-8 common objects, such as water bottles or hair-spray cans. 21 of these object sets were packed into boxes, and 2 were packed into bags together with background clutter such as clothing. The data was collected by physically rotating the bag or box inside the scanner, causing objects within to move slightly due to gravity. This relative motion is visually apparent in about 1/4 of our datasets. None of the data sets contained real explosives [3]. Figure 5(a) shows one collected bag, containing an electronic box, a water bottle and an umbrella.

We also manually annotated all scans by tracing each of the objects. These annotations were used to estimate the ground-truth effective atomic number and mass of each object. Three of these datasets serve as training data, and were used to estimate the scene-independent parameters: $\sigma_\ell^2 = 0.29, \gamma = 0.32, \sigma_A^2 = 100.2$. All results reported below were obtained on the remaining 20 datasets.

Figure 5(b) shows the decomposition obtained using **SATISϕ** across all views of this instance. Figures 5(c) and (d) show a similar decomposition for a single view. In all three datasets, the objects extracted using SATISϕ generally correspond to the true objects present in the bags, even though their images largely overlap in the scan.

For comparison, we considered two baseline methods. First, the **segmentation** baseline segmented the image using the method of [5] with no further processing. This method is state-of-the-art for extracting object-parts in a natural image. To the best of our knowledge, it is also the standard strategy in threat-detection applications. The parameters of the segmentation were tuned so as to be maximally effective in the task of automatic threat detection. Our second, **clustered** baseline, involved clustering the segments of the segmentation baseline into larger groups in an attempt to find full objects. X-Means [18] clustering was used in order to automatically select the number of clusters using an information criterion. Each segment was represented by a feature vector including the spatial centroid of the region and the estimated atomic number and density of the segment. Figure 4(b, 2^{nd} and 3^{rd} rows) shows decompositions according to our two baselines.

Figure 4(b) illustrates the difference between the operation of SATISϕ and the two baselines. The segmentation baseline attributes the entire attenuation of the circular segment to object #2, predicting an atomic number of 9.7 that corresponds to a nearly organic material. SATISϕ, on the other hand, correctly identifies that the vertical orange object forms a continuous rectangle, and that the attenuation in these pixels is due to both objects #2 and #3, giving the circular object an atomic number of 11.9, corresponding correctly to a light metal. The clustered baseline incorrectly merges these two objects into a single heavy organic object.

To quantitatively evaluate the performance of all methods, we further measured how accurately they identify the composition of substances in the scan sets. Given the separated objects, we converted their log-attenuation values into atomic number Z and mass M. Each object produces a single point in the two dimensional *MZ space* (mass – atomic number space). The quality of a decomposition is quantitatively evaluated by measuring the "distance" between the predicted Z-M point and the groundtruth one[4].

Figures 5(e,f) plot the predicted Z-M values for objects extracted by the two methods against the groundtruth values of the same objects. The plots correspond to the decompositions of data sets (b) and (c) respectively. In the example of (b), SATISϕ discovers an organic object (water bottle) beneath the electronic box. This object (with atomic number 5), matches very closely to the true object, as shown in the bottom left of plot (e). The segmentation approach, without the ability to attribute attenuations in that region to multiple objects, finds only high-Z (inorganic) objects.

Figure 6 quantifies these results for the entire collection of 20 datasets, containing 66 hand-annotated objects. For each hand-annotated object, we find the separated object with the highest spatial accuracy, as measured by overlap score. Overlap is measured by the number of pixels in the intersection of the true object and the separated object divided by the number of pixels in the union (producing a number between 0 and 1).

In the table of figure 6, we consider all matches with overlap at least 0.5. This results in 33 matches for **SATISϕ**, 31 for **segmentation**, and 29 for **clustered**. From these matches, we compute the error in estimating the material properties of the objects. The top row gives the root-mean-squared error (RMS) of the estimated mass as a proportion of the true mass. This error metric penalizes small absolute errors in small objects equally to larger absolute errors for large objects. The second row gives RMS error in atomic number (Z) estimate (atomic number ranges between 5-20 for most objects). On both of these metrics, SATISϕ significantly outperforms both baselines. The table also includes the total number of matched objects, the number of missed objects and the number of extra separated objects (separated

---

[3] The data set will be made available by email request.

[4] To make the comparison with the baseline segmentation more fair, and since it often produced tiny segments, we discarded objects whose mass was smaller than 2.5 per view.

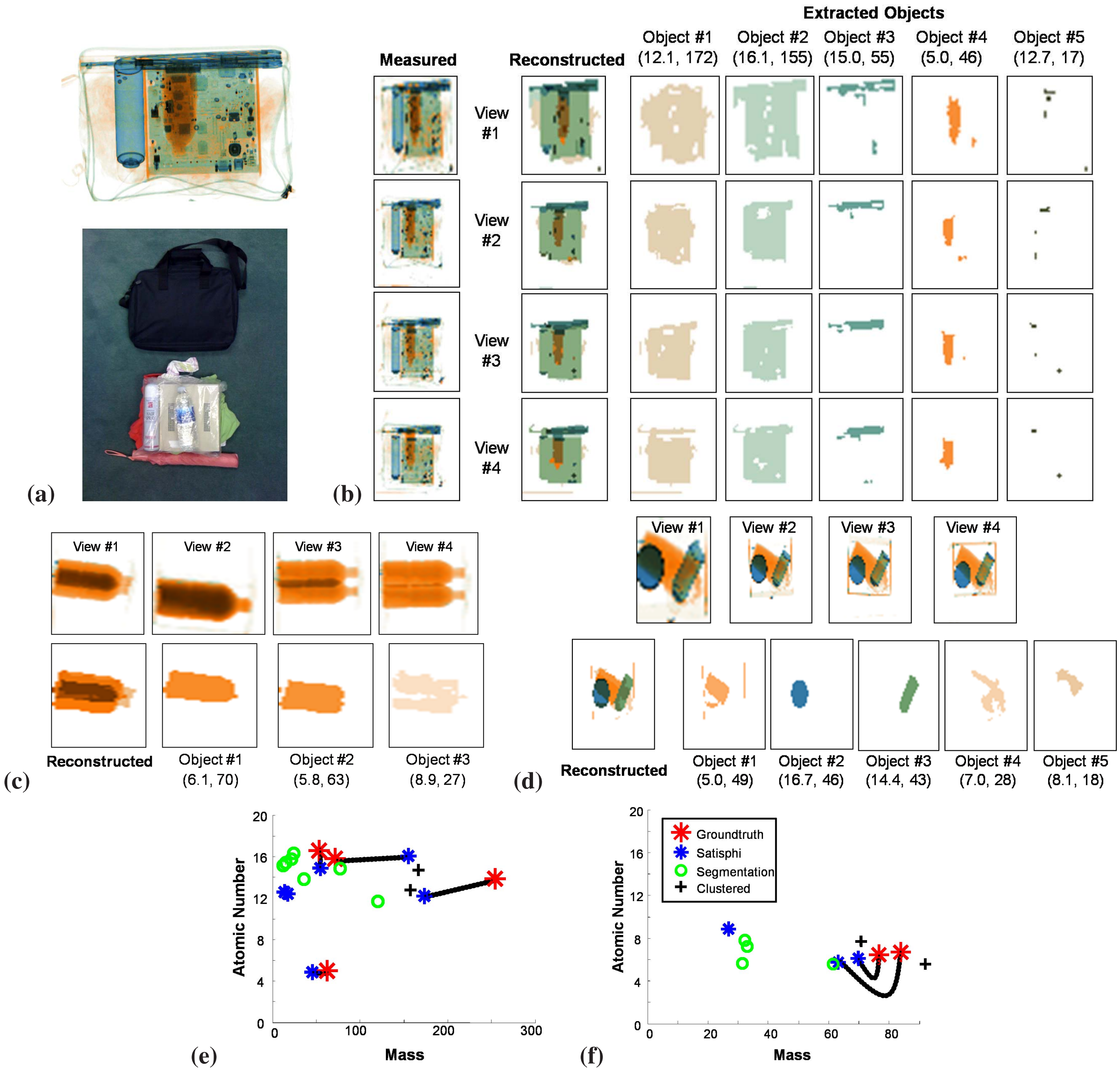


Figure 5. **(a)** One data instance, including an x-ray view (top) and a photograph of the bag and its content (bottom). **(b)** Reconstructions obtained by SATISϕ. The left column shows the original scans, the second column shows the reconstruction, and the 5 remaining columns correspond to 5 objects in the SATISϕ decomposition. Objects are colored based on material, and the text in parenthesis is the atomic number and mass (Z,M). **(c,d)** Results for two other scenes, showing the raw data (top row), and the reconstruction and objects for one of the views (bottom row). **(e,f)** Plots of the the objects discovered in (b) and (c) in MZ (mass – atomic number) space.

objects with no matching true object). This analysis shows that SATISϕ matches more objects correctly, with fewer extraneous objects than the baselines, while finding more accurate estimates of the material properties.

For another visualization of the quality of these matches, we consider the n "best" matches (those with highest overlap) for each method, and look at the RMS error in Z and relative mass for these matches. The plots to the right in figure 6 show the RMS error for Z and relative mass as a function of n (the number of matches included) for all methods. The left end of these curves represents the error over the easy matches(like isolated objects), which are often well estimated by all methods. Towards the middle of the graph, however, when the matches start to get more difficult, SATISϕ dramatically outperforms both baselines.

The primary error mode of SATISϕ appears to be oversplitting of true objects across layers. This is apparent in the example of figure 5(d), 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 better object approximations than the baselines.

	SATISϕ	Seg.	Clus.
RMS (mass)	0.66	1.15	1.32
RMS (Z)	1.55	2.52	2.71
Matched obj.	33	31	29
Missed obj.	33	35	37
Excessive obj.	22	61	12

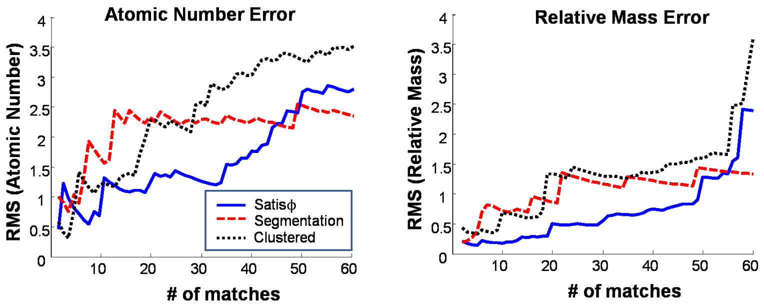


Figure 6. Quantitative evaluation over 20 datasets of the match between extracted and hand-annotated objects for our method and the two baselines. RMS is the average root-mean-squared error in either the error in relative mass or atomic number.

7. Discussion

This paper describes SATISϕ, a method that separates transmission images into overlapping objects that make up the image "layers.". It successfully disambiguates the objects with only a small number of views. We have described an efficient search method for finding a high-scoring decomposition, and have demonstrated the method's effectiveness on real x-ray scans. SATISϕ achieves an error reduction of 23% in the estimation of physical properties of objects over the baseline approach used in the automatic threat detection. Separating overlapping objects in x-ray scans has been long seen as a major barrier towards automatic threat detection, and our results may provide a basis for building more accurate detection systems.

CT, a standard x-ray approach used to compute an exact 3D reconstruction, requires thousands of costly scans and is limited to rigid non-moving objects. The probabilistic approach taken here allows us to extract the essential information that is needed for threat detection: 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 the introduction of 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 current 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 images with a small number of objects, and is therefore limited to low-clutter luggage. It remains a significant challenge to extend the results presented in this paper to highly cluttered scans. Furthermore, although it was designed to handle separating objects in x-ray images, the underlying ideas could apply to more general problems, such as segmenting semi-transparent objects [16] or reflected objects [13].

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. *Jrnl of Elec. 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, 8

[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 PAMI*, 24:1026–1038, 1999. 2

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

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *CVPR*, 2:264, 2003. 2

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

[8] V. Ganapathi, D. Vickrey, J. Duchi, and D. Koller. Constrained approximate maximum entropy learning. In *UAI*, 2008. 4

[9] S. Hata, Y. Saitoh, S. Kumamura, and K. Kaida. Shape extraction of transparent object using genetic algorithm. In *ICPR*, volume 4, page 684, 1996. 2

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

[11] 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

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

[13] A. Levin, A. Zomet, and Y. Weiss. Learning to perceive transparency from the statistics of natural scenes. *NIPS*, 2002. 2, 8

[14] D. Lide. *CRC handbook of chemistry and physics*. CRC press, 2004. 2, 3

[15] M. Lustig, D. Donoho, and J. M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, December 2007. 2

[16] K. McHenry, J. Ponce, and D. Forsyth. Finding glass. In *CVPR*, pages 973–979, 2005. 2, 8

[17] F. Metelli. The perception of transparency. *Scientific Amer.*, 230, 1974. 8

[18] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning*, pages 727–734, 2000. 6

[19] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and MRF MAP estimation. In *ICML*, pages 737–744, 2006. 5

[20] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, volume 1, pages 10–17, 2003. 5

[21] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, 2008. 2

[22] S. Soatto and P. Perona. Three dimensional transparent structure segmentation and multiple 3d motion estimation from monocular perspective image sequences. In *ICPR*, 1994. 2

[23] C. A. Sutton and A. McCallum. Piecewise training for undirected models. In *UAI*, pages 568–575, 2005. 4

[24] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. *CVPR*, 1:1246, 2000. 2

[25] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992. 2

[26] M. Yamazaki, S. Iwata, and G. Xu. Dense 3d recon. of specular and transparent objects using stereo cameras and phase-shift method. In *ACCV*, 2007. 2