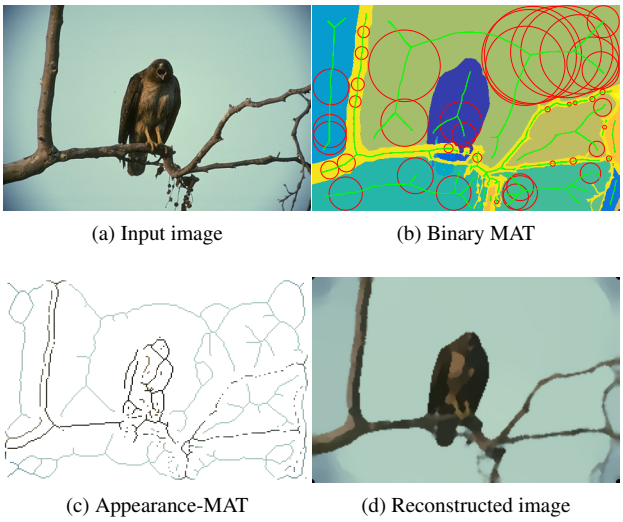
AMAT: Medial Axis Transform for Natural Images

Stavros Tsogkas, Sven Dickinson
University of Toronto
27 King's College Circle Toronto, Ontario M5S 1A1 Canada
{tsogkas,sven}@cs.toronto.edu

Abstract

We introduce Appearance-MAT (AMAT), a generalization of the medial axis transform for natural images, that is framed as a weighted geometric set cover problem. We make the following contributions: i) we extend previous medial point detection methods for color images, by associating each medial point with a local scale; ii) inspired by the invertibility property of the binary MAT, we also associate each medial point with a local encoding that allows us to invert the AMAT, reconstructing the input image; iii) we describe a clustering scheme that takes advantage of the additional scale and appearance information to group individual points into medial branches, providing a shape decomposition of the underlying image regions. In our experiments, we show state-of-the-art performance in medial point detection on Berkeley Medial AXes (BMAX500), a new dataset of medial axes based on the BSDS500 database, and good generalization on the SK506 and WH-SYMMAX datasets. We also measure the quality of reconstructed images from BMAX500, obtained by inverting their computed AMAT. Our approach delivers significantly better reconstruction quality w.r.t. three baselines, using just 10% of the image pixels. Our code and annotations are available at https://github.com/tsogkas/amat .



(a) Input image (b) Binary MAT

(c) Appearance-MAT (d) Reconstructed image

Figure 1: **Top:** Input image (1a) and segmentation (1b) from BSDS500, with color-coded ground-truth segments. Medial axes (green) and a subset of medial disks (red) are overlaid. Each (binary) segment can be reconstructed from its medial points and radii. **Bottom:** Similarly, the AMAT (1c) carries enough information to reconstruct the *input image* (1d) with just $\sim 5\%$ of the pixels.

1. Introduction

Symmetry is a ubiquitous property in the natural world, with a well-established role in human vision. Humans instinctively recognize and use symmetry to analyze complex scenes, as it facilitates the encoding of shapes and their discrimination and recall from memory [7, 34, 52]. In the context of computer vision, *local* symmetry is of particular interest, because of its robustness to viewpoint changes and its connection to salient structures, such as object parts. This intuition is fundamental to many milestones in object representation theory, including generalized cylinders [10], superquadrics [8], geons [9], and shock graphs [42].

Fundamental notions of local symmetry were introduced decades ago by Blum in the context of binary shapes with the *medial axis transform (MAT)* [11, 12]. The MAT is a powerful shape abstraction, and provides a compact representation that preserves topological properties of the input shape. These properties are invariant to translation, rotation, scaling, articulation, and their locality offers robustness to occlusion. The MAT has been very effective in reducing the computational complexity of algorithms for various tasks, including shape matching [42] and recognition [35], mesh editing [26, 55], and shape manipulation [15]. For these reasons many researchers have tried to achieve a good balance between MAT sparsity and reconstruction quality [46, 25].

Extending the notion of the MAT to natural images can correspondingly benefit applications that rely on a sparse

set of highly informative keypoints/landmarks, such as registration [56], retrieval [44, 5], pose estimation and body tracking [40], and structure from motion [2]. It could also assist segmentation by enforcing region-based constraints through their medial point representatives [48], and by providing a practical alternative to manual scribbles/seeds for interactive segmentation [13, 32, 20, 27]. Another interesting application is artistic rendering of images: [18] use approximate medial axes to simulate brush strokes and generate a painting-like version of the input photograph.

Unfortunately, the MAT has not found widespread use in tasks involving natural images, due to the lack of a generalization that accommodates color and texture. Previous works have mostly attacked *medial point detection* [49, 38], which amounts to determining the *locations* of points lying on medial axes but not the scale of the respective medial disks. The type of axes considered is also typically constrained to make the problem more concrete: [49] only considers elongated structures, on either foreground objects or background; [38] focuses on *object skeletons*, ignoring background structures. These methods lack another key characteristic of the MAT: medial point locations alone do not provide sufficient information to reconstruct the input.

In this paper we introduce the first "complete" MAT for natural images, dubbed *Appearance-MAT (AMAT)*. First, we provide a new definition in the context of natural images by framing MAT as a weighted geometric set cover (WGSC) problem. Our definition is centered around the MAT invertibility property and elicits a straightforward criterion for quality assessment, in terms of the reconstruction of the input image. Second, our algorithm associates each medial point with *scale* as well as local *appearance* information that can be used to reconstruct the input. Thus, the AMAT encompasses all the fundamental features of its binary counterpart. Third, we describe a simple bottom-up grouping scheme that exploits the additional scale and appearance information to connect points into medial *branches*. These branches correspond to meaningful image regions, and extracting them can support image segmentation and object proposal generation, while offering a shape decomposition of the underlying structure as well.

Being bottom-up in nature, our method does not assume object-level knowledge. It computes medial axes of both foreground and background structures, yielding a compact representation that only uses $\sim 10\%$ of the image pixels. Yet, this sparse set of points carries most of image signal, differing from other sparse image descriptions, *e.g*. edge maps, which strip the input of all appearance information.

We perform experiments in medial point detection on a new dataset of medial axes, the *Berkeley-Medial AXes (BMAX500)*, which is built on the popular BSDS500 dataset, showing state-of-the-art performance. We also measure the quality of reconstructions obtained by inverting the AMAT of images from the same dataset, using a variety of standard image quality metrics. We compare with three reconstruction baselines: one built on the medial point detection algorithm from [49] and two built from the ground-truth segmentations in BSDS500. Our method significantly outperforms the baselines in terms of reconstruction quality, while attaining a $11\times$ compression ratio.

The outline of the paper is as follows: we start by reviewing related work on medial axis extraction for binary shapes and natural images in Section 2. In Section 3 we describe our approach. Section 4 includes implementation details and in Section 5 we present our results. Finally, in Section 6 we conclude and discuss ideas for future directions.

2. Related Work

Binary shapes: Blum introduced the medial axis transform, or skeleton, of 2D shapes in his seminal works [11, 12]. Since then, researchers have developed algorithms for reliable and efficient medial axis extraction, its extension to 3D shapes, and its application to computer vision tasks.

Siddiqi *et al*. define *shocks* as the singularities of a curve evolution process acting on the boundaries of a shape, and they organize them into a directed, acyclic shock graph [42]. Shock graphs were successfully used in shape matching [42], recognition [35], and database indexing [36]. *Bone graphs* [29] offer improved stability and a more intuitive representation of an object's parts, by identifying and analyzing ligature structures. Visual part correspondences are also established and used to measure part and aggregated shape similarity in [22]. The correspondence of skeleton branches to object parts is further explored in [28, 6]. More recently, Stolpner *et al*. deal with the problem of approximating a 3D solid via a union of overlapping spheres [45].

The value of the MAT has been equally appreciated by the graphics community, where object shapes are routinely represented as point clouds or triangular meshes. Giesen *et al*. [17] introduced the *scale axis transform*, a skeletal shape representation that yields a hierarchy of successively simplified skeletons, which are obtained by multiplicative scaling of the MAT's radii. Li *et al*. [25] use quadratic error minimization to compute an accurate linear approximation of the MAT, called *Q-MAT*. They show experiments on medial axis simplification where they reduce the number of nodes of an initial medial mesh by three orders of magnitude, while preserving good surface reconstruction. A comprehensive compilation of medial methods and their applications in the binary setting can be found in [41].

Natural images: Compared to the binary setting, the number of works on medial axis detection for natural images is rather limited. Levinstein *et al*. [23] detect *symmetric parts* of objects by learning to merge adjacent deformable, maximally inscribed disks, modeled as superpix-

els. Learned attachment relations are then used to combine detected parts into coarse skeletal representations. Lee *et al*. extend that work by introducing a deformable disk model that can capture curved and tapered parts, and also add continuity constraints to the medial point grouping process [43]. In other works medial point detection is posed as a classification problem where pixels are labeled as "medial" or "not-medial", inspired by similar methods for boundary detection [30]. Tsogkas and Kokkinos use multiple instance learning (MIL) to deal with the unknown scale and orientation during training [49], while Shen *et al*. adapt a CNN with side outputs [53] for object skeleton extraction [38]. All these approaches exploit appearance information by incorporating a machine learning algorithm.

Our work can be regarded as lying at the intersection of previous work on binary and natural images. From a technical standpoint, it shares more similarities with binary methods, for instance [45], which solves the set cover problem for volumes in the 3D space. At the same time, it can be applied to real images, without assuming a figure-ground segmentation, but it also demonstrates unique characteristics. Our method does not involve learning, and is not constrained in detecting a particular subset of medial axes as [49, 38]. It also complements existing methods by augmenting point locations with scale and appearance descriptions, which are necessary for reconstructing the input.

3. AMAT definition

Consider a 2D binary shape, O, like the one in Figure 2, and its boundary Θ_O. The *medial axis* of O is the set of points $\mathbf{p}$ that are centers of the maximally inscribed (medial) disks, bitangent to Θ_O in the interior of the shape. The *medial* (disk) *radius* $r_{\mathbf{p}} \equiv r(\mathbf{p})$ is the distance between $\mathbf{p}$ and the points where the disk touches Θ_O. The process of mapping O to the set of pairs $(\mathbf{p}, r_{\mathbf{p}}) \in \mathbb{R}^2 \times \mathbb{R}$ is called the *medial axis transform* (MAT). Given these pairs, we can reconstruct O as a union of overlapping disks that sweep-out its interior by "expanding" a value of one (1) inside the area covered by each medial disk.

We argue that a MAT for real images should satisfy a similar principle: given the MAT of an image, we should be able to "invert" it, reconstructing *the image* itself. There are several reasons why extending this idea to real images is a challenging task: natural images depict complex scenes, cluttered with numerous objects, instead of just a single foreground shape. Moreover, unlike binary images, real images exhibit complicated color and texture distributions. Nevertheless, we can exploit image redundancies and assume that an image is composed of many small regions of relatively uniform appearance. This is the same assumption that underlies most superpixel algorithms which break up an image into non-overlapping patches, while respecting perceptually meaningful region boundaries [39, 24, 1].

Notation. In the rest of the paper we denote a disk of radius r, centered at point $\mathbf{p}$, as $D_{\mathbf{p},r} \equiv D(\mathbf{p}, r)$. For brevity, we often refer to such a disk as a r-disk or $(\mathbf{p}, r)$-disk. $\mathcal{D}$ is a collection of such disks of varying centers and radii, $\mathcal{D} = \{D_{\mathbf{p}_i, r_{\mathbf{p}_i}}\}$, $i \in \mathbb{N}$. The intersection of a $(\mathbf{p}, r)$-disk with an image I is a disk-shaped region of the image, and is denoted by $I \cap D_{\mathbf{p},r} = D^I_{\mathbf{p},r} \subset \mathcal{D}^{\mathcal{I}} = \{D^I_{\mathbf{p}_i, r_{\mathbf{p}_i}}\}$. Finally, we use $\circ$ to denote function composition, and $\|\cdot\|$ for an appropriate error metric (*e.g*., the L_2 norm).

Formulation. Consider an RGB image $I \subset \mathbb{R}^3$, and a disk-shaped region $D^I_{\mathbf{p},r} \subset I$. Let $f : \mathcal{D}^{\mathcal{I}} \to \mathbb{R}^K$ be a function that maps $D^I_{\mathbf{p},r}$ to a vector $\mathbf{f}_{\mathbf{p},r} = f \circ D^I_{\mathbf{p},r}$; we call $\mathbf{f}_{\mathbf{p},r}$ the *encoding* of $D^I_{\mathbf{p},r}$. Now let $g : \mathbb{R}^K \to \mathcal{D}^{\mathcal{I}}$ be a function that maps $\mathbf{f}_{\mathbf{p},r}$ back to a disk patch $\tilde{D}^I_{\mathbf{p},r} = \mathbf{g}_{\mathbf{p},r} = g \circ \mathbf{f}_{\mathbf{p},r}$. We call g the *decoding* function. In the general case, f and g will be *lossy* mappings, which means that the reconstruction error $e_{\mathbf{p},r} = \left\| \tilde{D}^I_{\mathbf{p},r} - D^I_{\mathbf{p},r} \right\| \geq 0$. Using the above, we define the AMAT as the set of tuples $M : \{(\mathbf{p}_1, r_{\mathbf{p}_1}, \mathbf{f}_{\mathbf{p}_1, r_{\mathbf{p}_1}}), \ldots, (\mathbf{p}_m, r_{\mathbf{p}_m}, \mathbf{f}_{\mathbf{p}_m, r_{\mathbf{p}_m}})\}$, such that:

$$M = \arg\min_{\mathbf{p},r} \sum_{i=1}^{m} e_{\mathbf{p}_i, r_i}, \quad I = \bigcup_{i=1}^{m} D^I_{\mathbf{p}_i, r_i}. \tag{1}$$

In Section 3.1 we discuss constraining m.

Encoding and decoding functions. Our framework allows f, g to take any form; for example, f could be a histogram representation of color in $D^I_{\mathbf{p},r}$ and g could return the mode of the distribution. In this paper we opt for simplicity: f computes the mean of each color channel "summarizing" $D^I_{\mathbf{p},r}$, in a 3×1 vector $\mathbf{f}_{\mathbf{p},r}$. Conversely, g constructs an approximation $\tilde{D}^I_{\mathbf{p},r} \approx D^I_{\mathbf{p},r}$ by replicating $\mathbf{f}_{\mathbf{p},r}$ in the respective disk-shaped area. When the $(\mathbf{p}, r)$-disk is fully enclosed in a uniform region the reconstruction error $e_{\mathbf{p},r}$ is low, whereas when the disk crosses a strong image boundary, the encoding $\mathbf{f}_{\mathbf{p},r}$ cannot accurately represent the underlying image region, resulting in a higher error.

Note that the definition in Equation (1) suggests conceptual similarities with superpixel representations. Selecting the points $\{(\mathbf{p}_i, r_i, \mathbf{f}_{\mathbf{p}_i, r_i})\}$, $i = 1, \ldots, m$ is equivalent to covering the input image with m disk-shaped superpixels. Minimizing the total reconstruction error implies that these "superdisks" do not cross region boundaries, as this would incur a high reconstruction error, as shown in Figure 2. However, there are two important differences: First, in our case a canonical shape (disk) is used, whereas superpixels can have any form. Second, our disks are *overlapping*, in contrast to standard, non-overlapping superpixels.

Using canonical shapes helps achieve sparsity of the final MAT. Disks are optimal in that sense, as they are rotationally invariant and are fully defined using only their

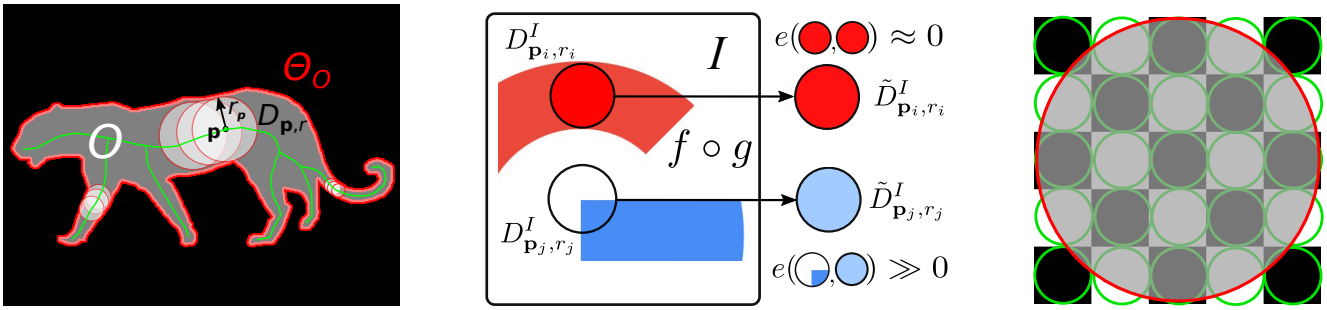


Figure 2: **Left:** We can reconstruct a binary shape by expanding a value of "1" within the area of all medial disks. **Middle:** Disks are represented by their mean RGB value; disks that cross region boundaries have a high reconstruction error. **Right:** Toy example: depending on the task, the user can favor a dense representation with low reconstruction error (green disks) or a sparse representation with high reconstruction error (red disk) by varying the scale parameter w_s.

center and radius. By contrast, a free-form element requires storing coordinates of *all* its boundary points. On the other hand, using one shape and no overlap would not reduce reconstruction quality, but it would result in disjointed medial points instead of smooth, connected medial axes.

3.1. AMAT as a Geometric Set Cover Problem

The geometric set cover is the extension of the well studied set cover problem, in a geometric space. Here we only consider the case of a two-dimensional space and we particularly focus on the *weighted* version of the problem, which is defined as follows: Consider a universe of N points $X \in \mathbb{R}^2$ and subsets $\mathcal{D} = \{D_1, D_2, \ldots, D_k\} \subseteq X$, called *ranges*. A common choice for D_i is intersections of X with simple shape primitives, such as disks or rectangles.

Now assume that each element in $\mathcal{D}$ is associated with a non-negative weight or *cost* c_i. Solving the WGSC problem amounts to finding a sub-collection $\bar{\mathcal{D}} \subset \mathcal{D}$ that covers the entire X (all N elements of X are contained in at least one set in $\bar{\mathcal{D}}$), while having the minimum total cost C; the total cost is simply the sum of costs of individual elements in $\bar{\mathcal{D}}$. WGSC is a strongly NP-hard problem for which polynomial-time approximate solutions (PTAS) exist. The interested reader can find more details on WGSC and related algorithms in [31, 50, 19, 14].

The AMAT formulation lends itself naturally to a WGSC interpretation. The spatial support X^I of an input image I, is the universe of N points. As $\mathcal{D}$ we consider the set of r-disks with r chosen from a finite set $\mathcal{R} : \{r_1, r_2, \ldots, r_R\}$. The r-disks can be placed at any position $\mathbf{p} = (x, y) \in X^I$ such that $D_{\mathbf{p},r}$ is fully contained in X^I. We also assign a cost $c_{ij} \equiv c_{\mathbf{p}_i,r_j} \propto e_{ij}$ to each $(\mathbf{p}_i, r_j)$-disk, $i \in [1, N]$, $j \in [1, R]$. Note that for brevity, we drop the subscripts $\mathbf{p}_i, r_j$ and simply use ij. We provide more details regarding computation of c_{ij} in Section 4.

As Equation (1) suggests, the goal is to find a subset of disks that cover the entire image, while maintaining a low total reconstruction cost. A trivial solution would be to select each pixel as a disk of radius $r = 1$, in which case $M = \{(\mathbf{p}_1, r_{\mathbf{p}_1}, f_{\mathbf{p}_1,r_{\mathbf{p}_1}}), \ldots, (\mathbf{p}_N, r_{\mathbf{p}_N}, f_{\mathbf{p}_N,r_{\mathbf{p}_N}})\}$, and $\sum_{i=1}^{N} e_{\mathbf{p}_i,r_i} = 0$; each pixel can be perfectly represented by its mean value. Such a solution is of no practical use. Staying true to the spirit of the MAT, we seek a solution that is *sparse* (low number of medial points m), while being able to adequately reconstruct the input image. One possible way to do this would be to agree on a fixed "budget" of points, and look for the optimal solution, given m. However, choosing an acceptable m can be a nuisance, as its value can vary significantly from image to image.

In the original MAT, sparsity is implicitly induced through the use of maximal disks, touching the shape boundary at two or more points. Extending the maximality principle to real images is not straightforward because color and texture boundaries are not robustly defined. Relying on the output of an edge extraction algorithm is not a viable option either, as it would make our method sensitive to errors from which it would be impossible to recover.

Instead, we choose to regularize the minimization criterion in Equation (1) by adding a scale-dependent term $s_j = \frac{w_s}{r_j} \propto \frac{1}{r_j}$ to the costs c_{ij}. This way we favor the selection of larger disks at each point, as long as s_j is not "too" large with respect to the error incurred by picking $D_{\mathbf{p},r_{j+1}}$ instead of $D_{\mathbf{p},r_j}$. Selecting a high value for w_s leads to a sparser solution with higher total reconstruction error, whereas a low value for w_s aims for a better reconstruction, by utilizing more, smaller disks to cover X^I. Figure 2 (right) shows a toy example of these two cases and Figure 3 shows how varying w_s progressively removes details in a real image, keeping only the coarser structures.

Greedy approximation algorithm. There are many polynomial-time-approximate-solution (PTAS) algorithms for the vanilla set cover problem and its geometric variants. Here we use the simple, greedy algorithm described in [51],

Algorithm 1 AMAT greedy algorithm.

Input: $X^I = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}, \mathcal{R} = \{r_1, \ldots, r_R\}, f, g$
Output: M

Initialization: $M \leftarrow \emptyset, X^c \leftarrow \emptyset$ ▷ X^c : covered pixels.
Compute $\mathbf{f}_{\mathbf{p},r}$, $\mathbf{g}_{\mathbf{p},r} = g \circ \mathbf{f}_{\mathbf{p},r}, c_{\mathbf{p},r}, \ \forall \mathbf{p} \in I, \forall r \in \mathcal{R}$
while $X^c \subset X^I$ **do**
  $c^e_{\mathbf{p},r} \leftarrow \frac{c_{\mathbf{p},r}}{|D_{\mathbf{p},r} \setminus X^c|} + \frac{w_s}{r}, \ \forall \mathbf{p} \in X^I, \ \forall r \in \mathcal{R}$
  $(\mathbf{p}^*, r^*) \leftarrow \arg\min_{\mathbf{p},r} c^e_{\mathbf{p},r}$,
  $c_{\mathbf{p},r} \leftarrow c_{\mathbf{p},r} - \frac{c_{\mathbf{p},r}}{|D_{\mathbf{p}^*,r^*} \setminus X^c|}$,
    $\forall \mathbf{p}, r : D_{\mathbf{p}^*,r^*} \cap D_{\mathbf{p},r} \neq \emptyset$
  $M \leftarrow M \cup (\mathbf{p}^*, r^*, f_{\mathbf{p}^*,r^*})$
  $X^c \leftarrow X^c \cup D_{\mathbf{p}^*,r^*}$
end while

adapted for the weighted case. The steps of our method are described in Algorithm 1. We start by computing the costs c_{ij} for all possible disks D_{ij}. We define the *effective cost* of D_{ij} as $c^e_{ij} = \frac{c_{ij}}{A_{ij}} + s_j$, where A_{ij} is the number of *new* pixels covered by D_{ij} (pixels that have not been covered by a previously selected disk). Starting from an empty set M, we pick the disk with the lowest c^e_{ij} and add it to the solution, removing the area D_{ij} from the remaining pixels to be covered. We also adjust the cost of all disks that intersect with D_{ij}, because each disk should be penalized only for the *new* pixels it is covering. This process is repeated until all image pixels have been covered by at least one disk.

3.2. Grouping Medial Points Into Branches

The scale and appearance associated with each medial point provide a rich description that can be used to group points belonging to the same region into *medial branches*. The beneficial effects of grouping in low-level vision tasks have been observed in previous works [16, 57, 21, 33]. In our case, grouping pixels into branches can help us refine the final medial axis, by aggregating consensus from neighboring points, and break the image into meaningful regions.

We group detected medial points using an agglomerative scheme that starts at fine scales and progressively merges together nearby points at coarser scales. Our grouping criterion relies on proximity in *scale-space* and *appearance*. Intuitively, points that lie close have higher probability of belonging to the same branch. We also expect that the scale of points will change *gradually* along a branch, so points that lie close to each other but have very different radii should probably not be grouped together. Finally, two points should not be grouped if their encodings are very dissimilar, regardless of their proximity in scale-space.

We initialize branches as the connected components of the AMAT output. Starting at a scale r_j, we consider one branch at a time, and examine all other branches within a neighborhood of size $r_j \times r_j$ and a scale neighborhood $[r_{j-3}, r_j]$. If two branches coexist in this scale-space neighborhood and their average encodings (summed along the branch curve) are similar, they are merged. The grouping algorithm terminates when all scales have been considered.

3.3. Medial Branch Simplification

The output of our algorithm captures mostly region centerlines but there are still imperfections in the form of noisy, disconnected medial point responses or "lumps", instead of thin contours. Such imperfections are expected because of the approximate solution to the minimization problem of Equation (1) and the use of a discrete grid.

Grouping MAT points into branches makes it possible to process each branch individually, enabling the correction of these errors post hoc. We perform simple morphological operations (dilation and thinning) on the points of each branch to merge neighboring and isolated pixels together, while removing redundant responses. We also adjust the scales of the medial points, to ensure that the medial disks corresponding to the simplified structure span the same image area. Because grouped branches correspond to relatively homogeneous regions, reconstruction results after simplification are practically identical. Examples of simplified medial axes are illustrated in Figure 4.

4. Implementation Details

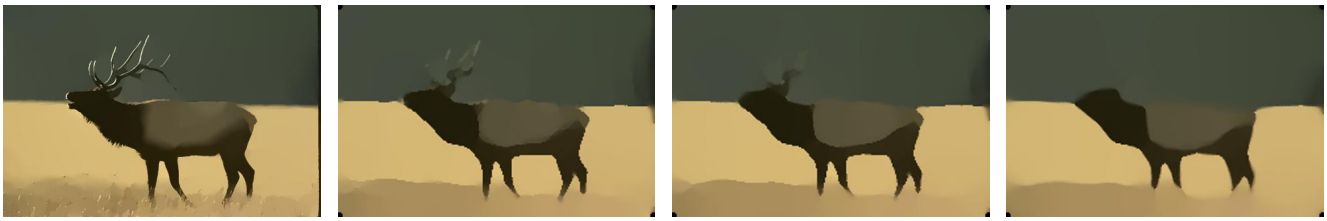
Disk Cost Computation. Using a simple error metric such as MSE to compute c_{ij} is not effective since disks with low MSE scores do not necessarily respect image boundaries. We propose the following alternative heuristic: First, we convert the RGB image to the CIELAB color space which is more suitable for measuring perceptual distances. Then, we define the cost of D_{ij} as

$$c_{ij} = \sum_k \sum_l \|\mathbf{f}_{ij} - \mathbf{f}_{kl}\|^2 \quad \forall k, l : D_{kl} \subset D_{ij}. \qquad (2)$$

Intuitively, a low cost c_{ij} implies that the encoding $\mathbf{f}_{ij}$ is representative of *all* disks that are fully contained in D_{ij}, hence D_{ij} is not crossing any region boundaries.

Dealing With Texture. The main motivation behind the choice of simple functions f, g, was simplicity and computational efficiency. Such functions also allow us to inject certain desired characteristics in the AMAT solution, such as appearance uniformity and alignment with boundaries.

However, natural images often contain high-frequency textures or noise, which can lead to the accumulation of large errors in Equation (2), and promote the selection of disks that do not correspond to perceptually coherent regions. Simple processing techniques (*e.g.*, Gaussian filtering) can reduce noise but they also degrade image boundaries and blend together neighboring regions.



(a) Input image (b) $w_s = 10^{-4}$ (c) $w_s = 10^{-3}$ (d) $w_s = 10^{-2}$

Figure 3: Using a progressively larger scale-cost factor w_s removes details, keeping only coarse image structures.

To alleviate this problem, we "simplify" the input image before extracting the AMAT, using a method that smooths high frequency regions, while preserving important edges [54]. In practice, this preprocessing produces an image that is perceptually very similar to the original, but without high-frequency textures that can cause the greedy algorithm to fail by placing disks at undesired locations.

Inverting the AMAT. Generating the reconstruction of a single disk-shaped region, $\tilde{D}^I_{\mathbf{p},r}$, is trivially achieved by replicating $\mathbf{f}_{\mathbf{p},r}$. However, since medial disks overlap, most pixels in the image domain will be covered by multiple disks with different encodings. We resolve this ambiguity in a simple way: while computing the AMAT, we keep track of the number of disks each pixel is covered by; this quantity is called *depth* in the context of the set cover problem. We then use the average $\mathbf{f}$ of all disks covering a point $\mathbf{p}_i$ with depth d_i as its reconstructed value:

$$\tilde{I}(\mathbf{p}_i) = \frac{1}{d_{\mathbf{p}_i}} \sum_{\mathbf{p},r} \mathbf{f}_{\mathbf{p},r}, \quad \forall \mathbf{p}, r : \mathbf{p}_i \in D_{\mathbf{p},r}. \tag{3}$$

Parameter Values. For the smoothing algorithm we use the default values $\lambda = 2 \cdot 10^{-4}$ and $\kappa = 2$ that the authors suggest for natural images [54]. Regarding the scale cost term described in Section 3.1, we found that $w_s = 10^{-4}$ is a value that strikes a good balance between reconstruction quality and sparsity of the generated medial axis. The maximum radius R must be finite to keep complexity manageable, but large enough to capture large uniform structures in the image. Based on the size of images used in our experiments we used 40 scales, excluding $r = 1$ to force disks to be larger than single pixels; thus $r \in [2, 41]$.

Complexity and Running Time. Computing c_{ij} requires computing differences for all disks in D_{ij}. If r_j is large, this number can grow quickly, yielding $O(NR^4)$ complexity. However, the most time-consuming step is the greedy approximation algorithm: At each iteration we cover at most $O(R^2)$ pixels, but we also have to update the costs of all overlapping disks. This has $O(NR^2 \sum_{r=1}^{R} r^2) = O(NR^5)$ complexity. One could parallelize the procedure by partitioning an image, simultaneously processing individual parts, and combining the results. Our single-thread MATLAB implementation takes $\sim 30\,\text{sec}$ for the AMAT, grouping, and simplification steps, on a 256×256 image.

5. Experiments

We evaluate the performance of our method on two tasks: i) localization of medial points in an image; and ii) generating accurate reconstructions of images, given their AMAT.

5.1. Medial Point Detection

We want to emphasize the difference between the problem we are addressing and the objectives pursued in other works. In [49] the authors focus on detecting local reflective symmetries of elongated structures, and they build a dataset with annotations of segments in the BSDS300 that fit this description. As a result, a large portion of the segments in BSDS is not used in performance evaluation. In [38] the authors are explicitly interested in extracting *object* skeletons, completely ignoring background structures. Although extracting object skeletons may be convenient for some tasks, it does not constitute a generalized notion of MAT.

In our work we do not make such distinctions. The central idea behind the AMAT is to be able to reproduce the full input image, so we view all parts of the image as equally important. This is also the reason we choose BSDS500 as a basis for constructing medial axes annotations. BSDS500 contains multiple segmentations for each image, offering higher probability of capturing segments at varying scales, making it more relevant to the problem we are trying to solve than datasets with object-level annotations.

Following [49], we individually apply a skeletonization algorithm [47] to binary masks of all segments in a given segmentation, extracting *segment skeletons*. The medial axis ground-truth for the image is formed by taking the union of all the segment skeletons, and this process is repeated for all available annotations (usually 5-7 per image). To conduct a fair comparison, we retrain the CG+BG+TG variant (MIL-color) from [49] on BMAX500. We also tried to retrain the CNN used in [38], but the outputs we obtained

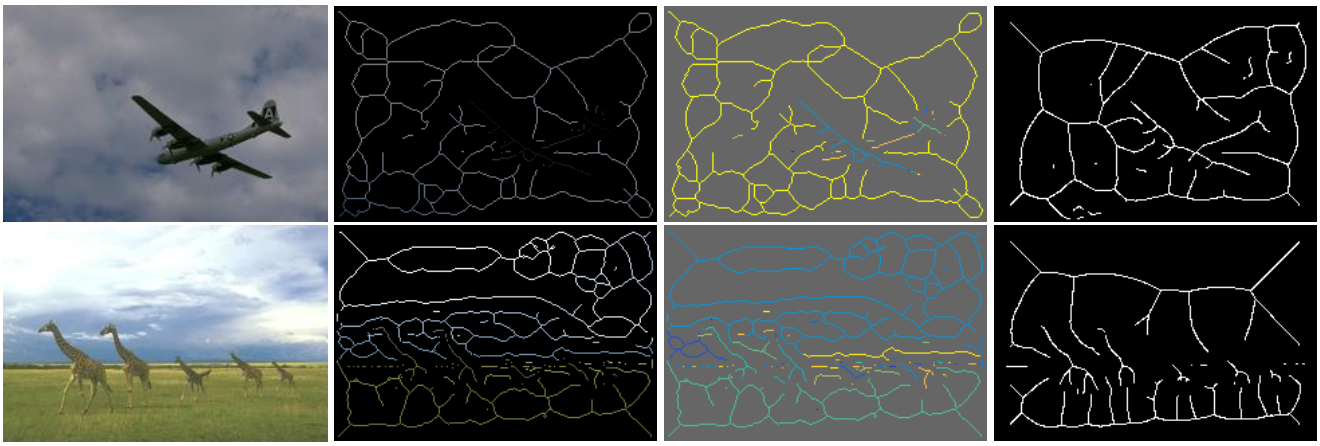


Figure 4: From left to right: Input image, AMAT axes (unused points in black), medial point groups (color-coded), ground-truth skeletons. Note that semantically coherent image regions (*e.g.*, sky, grass) tend to be grouped together.

were too noisy, and of no practical use. We hypothesize that this is because of the lack of consensus among the multiple ground-truth maps available for each image, which leads to convergence problems for the network; this has been previously reported in [53]. We evaluate performance using the standard precision, recall and F-measure metrics, and show the superior results of our method in Table 1. Note that our algorithm outputs binary skeletons, so plotting a PR-curve by varying a score threshold is not applicable in our case. "Human" performance is defined in the same manner as in [30, 49]. For all methods, detections within a distance of 1% of the image diagonal from a ground-truth positive are considered as true positives. We show qualitative results of the medial axes and the grouped branches in Figure 4.

Segmentation + skeletonization: As an additional baseline we compute skeletons after running Arbelaez's segmentation algorithm [3, 4] at scales 0.2 (F=0.61), 0.3 (F=0.58), 0.4 (F=0.54), 0.5 (F=0.5). We point out that the performance of UCM + skeletonization depends critically on the threshold selection. The optimal threshold is not known a-priori and, given a desired level of skeleton detail, the appropriate value varies from image to image. By contrast, AMAT's scale parameter is more intuitive to select and provides image-independent control of skeleton detail.

SK506 and WH-SYMMAX: We also evaluate the performance of the AMAT on two additional datasets: WH-SYMMAX [37] (F=0.44) and SK506 [38] (F=0.33). We compare with the pretrained FSDS [38] evaluating only on foreground skeletons, since our approach does not distinguish foreground from background. FSDS performs better than AMAT (F=0.67 and F=0.45 respectively). This is unsurprising, given that FSDS is a supervised method trained on these datasets in a way that allows it to take advantage of rich, object-specific information. However, this specialization comes at a cost: FSDS cannot generalize well to structures it has not seen before, which is evident when running it on BMAX500 (F=0.34 *vs*. F=0.56 for AMAT).

Metric	Precision	Recall	F-measure
MIL [49]	0.49	0.55	0.52
AMAT	0.52	0.63	**0.57**
Human	0.89	0.66	0.77

Table 1: Medial point detection on the BSDS500 val set.

5.2. Image Reconstruction

We now assess the quality of reconstructions we obtain by inverting the computed AMAT of images from the BSDS500 dataset. We compare with a baseline reconstruction algorithm based on the MIL approach of [49] (after retraining MIL-color on BMAX500). Their method uses features extracted in rectangular areas to produce a map of medial point strength at 13 scales and 8 orientations, for each pixel. A single confidence value for each point is derived through a noisy-or operation, which does away with scale and orientation information. As a surrogate, in our experiments we associate each point with the scale/orientation combination that has the highest score.

The scheme we use to create a crude reconstruction with their approach is the following: We start by sorting medial point scores in decreasing order and we pick the highest-scoring point. The rectangular region at the respective scale and orientation is then marked as covered, and the process is repeated until the whole image has been reconstructed. Similarly to our own method, point encodings are the mean RGB values within the rectangle, and local reconstructions are computed by averaging overlapping encodings. We also

Figure 5: **Image reconstruction**. From left to right: Input image, MIL [49], GT-seg, GT-skel, AMAT.

Metric	MSE	PSNR (dB)	SSIM	Compression
MIL [49]	0.0258	16.6	0.53	20×
GT seg	0.0149	18.87	0.64	9×
GT skel	0.0114	20.19	0.67	14×
AMAT	**0.0058**	**22.74**	**0.74**	11×

Table 2: Image reconstruction quality in BSDS500 val set.

compare with two more baselines: one obtained by considering ground-truth (GT) segments in BSDS500 and representing them by their mean RGB values (GT-seg); and a second, obtained through the GT skeletons and radii in BMAX500 (GT-skel). For the latter, we use the reconstruction process described in Section 4.

We consider three standard evaluation metrics for image similarity: MSE, PSNR, and SSIM. Results are reported in Table 2 and visual examples are shown in Figure 5. MIL uses rectangle filters at a finite set of scales and orientations that do not always match the scale and orientation of structures present in an image. As a result, MIL reconstructions are very blurred. GT-based reconstructions, on the other hand, have sharp edges but tend to have less texture detail, because people tend to undersegment images, favoring *perceptual* coherence over region appearance coherence. Note that, for each image, we choose the GT annotation that produces the best SSIM score, to ensure we are always comparing against the best possible GT-based reconstruction.

6. Discussion

We have defined the first complete medial axis transform for natural images. Our approach bridges the gap between MAT methods for binary shapes and medial axis/local symmetry detection methods for real images. We have demonstrated state-of-the-art performance in medial point detection and shown that we can produce a high-quality rendering of the input image using as few as 10% of its pixels.

That said, it is important to note that AMAT is not designed to be optimal for either of these tasks. Instead, it is designed to strike a balance between two conflicting objectives: i) capturing an image's salient structures (in the form of medial axes and their respective scale/appearance information); ii) providing an accurate reconstruction of the original image from this abstracted representation. Therefore, performance should be assessed on both objectives jointly.

We also want to emphasize that AMAT is a purely bottom-up algorithm, completely unsupervised and train-free. We consider this an important advantage of our approach, as it means that it can generalize well and in a predictable way to new datasets, without the need for additional tuning. Despite the lack of training, we have shown that it performs surprisingly well, and can even be competitive with supervised methods fine-tuned to specific datasets.

In future work, our goal is to parameterize our method to accommodate the relative roles of shape and appearance, and allow for flexible hierarchical grouping of medial branches to support segmentations of varying granularities. Furthermore, although our current choice of f/g favors simplicity and compactness at the cost of texture, our framework can accommodate *any* encoding/decoding functions. Designing alternatives to better capture and reconstruct texture, or for specific discriminative tasks, is another exciting future direction.

Acknowledgements

This work was funded by NSERC. We thank Ioannis Gkioulekas for his valuable suggestions and feedback.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Suesstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012. 3

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 2

[3] P. Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *Workshop on Perceptual Organization in Computer Vision (POCV)*, 2006. 7

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011. 7

[5] Y. Avrithis and K. Rapantzikos. The medial feature detector: Stable regions from image boundaries. In *ICCV*, 2011. 2

[6] X. Bai and L. J. Latecki. Path similarity skeleton graph matching. *TPAMI*, 2008. 2

[7] H. Barlow and B. Reeves. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision research*, 1979. 1

[8] A. H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1981. 1

[9] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 1987. 1

[10] T. O. Binford. Visual perception by computer. In *IEEE conference on Systems and Control*, 1971. 1

[11] H. Blum. A transformation for extracting new descriptors of shape. *Models for the perception of speech and visual form*, 1967. 1, 2

[12] H. Blum. Biological shape and visual science (part i). *Journal of theoretical Biology*, 1973. 1, 2

[13] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *ICCV*, 2001. 2

[14] T. M. Chan, E. Grant, J. Könemann, and M. Sharpe. Weighted capacitated, priority, and geometric set cover via improved quasi-uniform sampling. In *SODA*, 2012. 4

[15] H. Du and H. Qin. Medial axis extraction and shape manipulation of solid objects using parabolic pdes. In *Symposium on Solid modeling and applications*, 2004. 1

[16] P. Felzenszwalb and D. McAllester. A min-cover approach for finding salient curves. In *CVPRW*, 2006. 5

[17] J. Giesen, B. Miklos, M. Pauly, and C. Wormser. The scale axis transform. In *Symposium on Computational geometry*, 2009. 2

[18] B. Gooch, G. Coombe, and P. Shirley. Artistic vision: painterly rendering using computer vision techniques. In *International Symposium on Non-photorealistic animation and rendering*, 2002. 2

[19] S. Har-Peled and M. Lee. Weighted geometric set cover problems revisited. *Journal of Computational Geometry*, 2012. 4

[20] H. Isack, O. Veksler, M. Sonka, and Y. Boykov. Hedgehog shape priors for multi-object segmentation. In *CVPR*, 2016. 2

[21] I. Kokkinos. Highly accurate boundary detection and grouping. *CVPR*, 2010. 5

[22] L. J. Latecki and R. Lakamper. Shape similarity measure based on correspondence of visual parts. *TPAMI*, 2000. 2

[23] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Multiscale symmetric part detection and grouping. *IJCV*, 2013. 2

[24] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi. Turbopixels: Fast superpixels using geometric flows. *TPAMI*, 2009. 3

[25] P. Li, B. Wang, F. Sun, X. Guo, C. Zhang, and W. Wang. Q-mat: Computing medial axis transform by quadratic error minimization. *TOG*, 2015. 1, 2

[26] X. Li, T. W. Woon, T. S. Tan, and Z. Huang. Decomposing polygon meshes for interactive applications. In *Symposium on Interactive 3D graphics*, 2001. 1

[27] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016. 2

[28] H. Ling and D. W. Jacobs. Shape classification using the inner-distance. *TPAMI*, 2007. 2

[29] D. Macrini, S. Dickinson, D. Fleet, and K. Siddiqi. Bone graphs: Medial shape parsing and abstraction. In *CVIU*, 2011. 2

[30] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *TPAMI*, 2004. 3, 7

[31] N. H. Mustafa, R. Raman, and S. Ray. Quasi-polynomial time approximation scheme for weighted geometric set cover on pseudodisks and halfspaces. *SICOMP*, 2015. 4

[32] B. L. Price, B. Morse, and S. Cohen. Geodesic graph cut for interactive image segmentation. In *CVPR*, 2010. 2

[33] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo. Making better use of edges via perceptual grouping. In *CVPR*, 2015. 5

[34] F. L. Royer. Detection of symmetry. *Journal of Experimental Psychology: Human Perception and Performance*, 1981. 1

[35] T. Sebastian, P. Klein, and B. Kimia. Recognition of shapes by editing shock graphs. In *ICCV*, 2001. 1, 2

[36] T. B. Sebastian, P. N. Klein, and B. B. Kimia. Shock-based indexing into large shape databases. In *ECCV*, 2002. 2

[37] W. Shen, X. Bai, Z. Hu, and Z. Zhang. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images. *Pattern Recognition*, 2016. 7

[38] W. Shen, K. Zhao, Y. Jiang, Y. Wang, Z. Zhang, and X. Bai. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In *CVPR*, 2016. 2, 3, 6, 7

[39] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000. 3

[40] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al. Efficient human pose estimation from single depth images. *TPAMI*, 2013. 2

[41] K. Siddiqi and S. Pizer. *Medial representations: mathematics, algorithms and applications*. Springer Science & Business Media, 2008. 2

[42] K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker. Shock graphs and shape matching. *IJCV*, 1999. 1, 2

[43] T. Sie Ho Lee, S. Fidler, and S. Dickinson. Detecting curved symmetric parts using a deformable disc model. In *ICCV*, 2013. 3

[44] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[45] S. Stolpner, P. Kry, and K. Siddiqi. Medial spheres for shape approximation. *TPAMI*, 2012. 2, 3

[46] R. Tam and W. Heidrich. Shape simplification based on the medial axis transform. In *VIS*, 2003. 1

[47] A. Telea and J. Van Wijk. An augmented fast marching method for computing skeletons and centerlines. *Eurographics*, 2002. 6

[48] C. L. Teo, C. Fermueller, and Y. Aloimonos. Detection and segmentation of 2d curved reflection symmetric structures. In *ICCV*, 2015. 2

[49] S. Tsogkas and I. Kokkinos. Learning-based symmetry detection in natural images. In *ECCV*, 2012. 2, 3, 6, 7, 8

[50] K. Varadarajan. Weighted geometric set cover via quasi-uniform sampling. In *STOC*, 2010. 4

[51] V. V. Vazirani. *Approximation algorithms*. 2013. 4

[52] J. Wagemans. Parallel visual processes in symmetry perception: Normality and pathology. *Documenta ophthalmologica*, 1998. 1

[53] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015. 3, 7

[54] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via l_0 gradient minimization. In *TOG*, 2011. 6

[55] S. Yoshizawa, A. G. Belyaev, and H.-P. Seidel. Free-form skeleton-driven mesh deformations. In *Symposium on Solid modeling and applications*, 2003. 1

[56] Y. Zhou, E. Antonakos, J. Alabort-i Medina, A. Roussos, and S. Zafeiriou. Estimating correspondences of deformable objects. In *CVPR*, 2016. 2

[57] Q. Zhu, G. Song, and J. Shi. Untangling cycles for contour grouping. In *ICCV*, 2007. 5