Appearance-MAT: Extending the Medial Axis Transform to Color Images

Stavros Tsogkas, University of Toronto

November 1, 2016

1 Medial axis transform for binary shapes

Blum defined the Medial Axis of a shape, as the set of points that lie in the interior of the shape and that are equidistant from its boundary. An alternative way of defining the medial axis is to imagine all possible circles that are entirely contained in the interior of the shape, and touch the shape boundary at *exactly two points*. If we connect the centres of all these circles, we obtain the medial axis of the shape, which is alternatively called the *skeleton* of the shape.

In addition to the position, the medial axis definition also involves a *scale* parameter at each point. Scale represents the distance of the skeleton from the boundary or, equivalently, the radius of the circle that is centred around the point and touches the boundary. The procedure of converting the boundary of a shape to its medial axis is called the *medial axis transform*, or simply *MAT*. The inverse transformation is also possible: given the MAT of shape, we can *precisely* reconstruct the boundaries of the input, thus reconstructing the shape itself.

2 Extracting medial axes from natural images

Extending the above definition in the context of natural images would suggest the following two-step procedure: first, extract the boundary of a shape; second, apply the MAT to this boundary. However, despite significant progress in computer vision, reliably extracting object boundaries remains challenging when dealing with natural images. Even when a crude boundary can be extracted, tiny inaccuracies can drastically change the corresponding medial axis, making this approach non-robust and practically inapplicable. What we would really like to do is have a way to directly infer the medial axis from raw image data, without relying on an intermediate boundary extraction step.

Tsogkas and Kokkinos took first steps in developing such an algorithm for directly extracting skeletons, using a data-driven learning approach [**?**]. Learning approaches have proven very fruitful in the extraction of other low- and mid-level features, such as boundaries [**?**, **?**], corners [**?**] and junctions [**?**, **?**]. Their approach reduces medial point extraction to a supervised classification problem: features tailored to capture local symmetry at multiple scales and orientations, are extracted in regions centred around each point in the image. Then they use the multiple instance learning (MIL) framework [**?**, **?**] to train a classifier that computes the probability of each pixel belonging to an object's medial axis, at *any* combination of scale and orientation. Shen et al use the more powerful machine learning machinery of convolutional neural networks (CNN), for the same task, significantly improving medial point detection performance [**?**].

Both methods take a colour image as input and produce a dense probability map of "medialness" without requiring a boundary extraction step. However, their output does not constitute a complete MAT, as it is characterized by some important limitations. First, they only provide locations of medial axis points, without any scale information; remember that scale is an essential component of the binary MAT, since it allows us to reconstruct the original shape. Local scale estimation can be achieved in [**?**, **?**] by computing symmetry responses for a finite set of scales and then simply select the scale that gives the highest response.

Unfortunately this is not ideal, as a particular point can exhibit symmetry at more than one scales. For example, the center of a rectangle lies in three axes of reflective symmetry, at $0°, 90°$, and $180°$. Handling such ambiguities is challenging. Should we associate every point with a single scale? And if yes, using what criteria?

A second, fundamental limitation of these methods is that, given their data-driven nature, they rely on ground truth data for the supervised training of the algorithms.

3 AMAT definition

In order to extend the notion of the MAT to natural colour images, while avoiding the pitfalls analyzed in Section 2, we ground our approach on two key components of the original MAT for binary images: a) associating each medial point with a particular scale; b) being able to reconstruct the input, given its MAT. The latter will also provide us with a suitable criterion for evaluating the accuracy of our algorithm, which we will call *Appearance-MAT*, or *AMAT*, because it relies on complex appearance information (colour and/or texture) rather that 0/1 pixel activations. More specifically, we assess the quality of an AMAT by its capacity of reconstructing the original input. If $A(I)$ is the function that extracts the MAT from a color image I, and A^{-1} is its inverse function, then AMAT can be equivalently described by the following minimization:

$$AMAT(I) \equiv \min \left(d \left(A^{-1} \left(A(I) \right) - I \right) \right), \tag{1}$$

where $d(\cdot)$ is a distance function, defined on the image domain.

We will now define in detail the notation and quantities that we use in the AMAT formulation. We start by considering a RGB image I, defined over a domain $\mathcal{I} \subset \mathbb{R}^2$, and assuming $D(\mathbf{p}_c, r)$ to be the circular disk of radius r, centred at pixel $\mathbf{p}_c = (x_c, y_c) \in \mathcal{I}$:

$$D(\mathbf{p}_c, r) = \{\mathbf{p} : \|\mathbf{p} - \mathbf{p}_c\| \leq r\}, \quad \mathbf{p} = (x, y) \in \mathcal{I}, \quad r \in \mathbb{R}. \tag{2}$$

Next we will define two functions; the first one is a function that maps a disk region of RGB image values to a co-domain $\mathcal{Y}$:

$$f\left(I\left(D(\mathbf{p}, r)\right)\right) = f\left(D_I(\mathbf{p}, r)\right) : \mathbb{R}^3 \mapsto \mathcal{Y}. \tag{3}$$

$\mathcal{Y}$ depends on the choice of f; for example, if f computes the average of the image pixels within the circular disk, $\mathcal{Y}$ is a subset of $\mathbb{R}$. Alternatively, if f computes a histogram of RGB values in $D(\mathbf{p}, r)$, $\mathcal{Y}$ is a subset of $\mathbb{R}^b$, where b is the number of bins used to assign the values of each color channel. We have also used D_I as an abbreviation for the set of image values (RGB triplets) extracted from the circular disk region D.

The second function we need to define is a function that maps a pair of point coordinates (x, y) and an associated scalar r to a disk patch of RGB values, of radius r, centred at point $\mathbf{p} = (x, y)$:

$$g(\mathbf{p}, r) = \bar{D}_I(\mathbf{p}, r) : \mathcal{I} \times \mathbb{R} \mapsto \mathbb{R}^3 \times \mathcal{D}_r. \tag{4}$$

Here $\mathcal{D}_r$ is the set of all disks of radius r. It is important to stress that, although g can be viewed as performing the inverse operation of f, $g \neq f^{-1}$ in the strict analytical sense. In other words, $\bar{D}_I$ is generally an approximation of the original disk patch. Using $\circ$ to denote function synthesis. $g \circ f \circ D_I(\mathbf{p}, r) = \bar{D}_I(\mathbf{p}, r) \approx D_I(\mathbf{p}, r)$. If $\bar{D}_I(\mathbf{p}, r) = D_I(\mathbf{p}, r),\ \forall (\mathbf{p}, r) \in \mathcal{I} \times \mathbb{R}$ then $g \equiv f^{-1}$.

At this point we have defined all the tools that are needed to codify AMAT. The first step of our algorithm is to select appropriate mappings f, g, as well as a distance function d that will be used to evaluate performance, according to the criterion in Equation 1. The second step is to compute $f\left(D_I(\mathbf{p}, r)\right),\ \forall (\mathbf{p}, r) \in \mathcal{I} \times \mathcal{R}$, where $\mathcal{R} \subset \mathbb{R}$ is a finite set of discrete radii values. After computing f at all possible locations in the image and at all chosen scales (each radius corresponds to a different scale of the pixel neighbourhood), we

want to select a subset $M = \{(\mathbf{p}_1, r_1), (\mathbf{p}_2, r_2), \ldots, (\mathbf{p}_n, r_n)\}$, of n points and respective radii, such that:

$$\bigcup_{i=1}^{n} D(\mathbf{p}_i, r_i) = \mathcal{I} \tag{5}$$

$$M = \arg\min_{\mathbf{p},r} d\left(\bigcup_{i=1}^{n} \bar{D}_I(\mathbf{p}_i, r_i), I\right) \tag{6}$$

Essentially, our goal is to cover the entire image domain, while generating a reconstruction that is as close to the original image as possible. A caveat is that one could simply select a very large number of small disks that locally reconstruct the image with high accuracy. This approach would yield an output that is very close to the input but it would also result in a meaningless medial axis representation, as n would be close to the number of pixels in the image.

A way to circumvent this is to enforce additional constraints on the choice of points in M, by incorporating the notion of the *maximal disk*, used in Blum's original work [**?**]. To do that we consider a second metric, e, which we use to measure the distance between the encodings $f\left(D_I(\mathbf{p}_i, r_i)\right), f\left(D_I(\mathbf{p}_j, r_j)\right)$ of two disk patches, and a threshold ϵ. If $e\left(f\left(D_I(\mathbf{p}_i, r_i)\right), f\left(D_I(\mathbf{p}_j, r_j)\right)\right) < \epsilon$ and $r_i < r_j$, the larger disk "subsumes" the smaller one, excluding $(\mathbf{p}_i, r_i)$ from M, ensuring that a homogeneous image neighbourhood is represented (covered) by the largest disk possible (maximal disk). Algorithm 1 summarizes the above steps in pseudocode form.

Algorithm 1 Pseudocode for the Appearance-MAT algorithm.

procedure AMAT($I, e, d, f, g, \epsilon, \mathcal{R}$) ▷ Image I has N pixels and domain $\mathcal{I}$.
  Initialize solution as $M_0 = \{(\mathbf{p}_1, r_1), \ldots, (\mathbf{p}_N, r_N)\}$.
  Compute $f\left(D_I(\mathbf{p}, r)\right),\ \forall D_I(\mathbf{p}, r) \in I$.
  for all $\mathbf{p}_i, \mathbf{p}_j \in \mathcal{I},\ r_i, r_j \in \mathcal{R}$ **do** ▷ Remove non-maximal disks.
    if $e\left(f\left(D_I(\mathbf{p}_i, r_i)\right), f\left(D_I(\mathbf{p}_j, r_j)\right)\right) < \epsilon$ and $r_i < r_j$ **then** $M_0 \leftarrow M_0 - \{(\mathbf{p}_i, r_i)\}$.
    end if
  end for
  $M = \arg\min_{(\mathbf{p},r)\in M_0} d\left(\bigcup_{i=1}^{n} \bar{D}_I(\mathbf{p}_i, r_i), I\right)$ ▷ Select subset that optimizes reconstruction quality.
end procedure

4 Minimizing the reconstruction criterion

Solving Equation 6 depends on the choice of the encoder/decoder functions f, g, the distance functions e, d, the threshold ϵ, and the range of radii $\mathcal{R}$. A smart choice of these functions can significantly decrease the complexity of our algorithm, while improving reconstruction quality. The assumption that we make is that the distance functions e, d are additive. In the context of $AMAT$, this means that, given two RGB patches D_1, D_2 and and image I, the following property holds:

$$\begin{aligned} d(D_1 \cup\ D_2, I) &= d(D_1 \cup\ D_2, I \cap (D_1 \cup\ D_2)) \\ &= d(D_1, I \cap D_1) + d(D_2, I \cap D_2) - d(D_1 \cap\ D_2, I \cap (D_1 \cap\ D_2)) \end{aligned} \tag{7}$$

The above property will allows us to use dynamic programming to obtain the optimal solution M. First, let D^n be the union of n disk patches: $D^n = \bigcup_{i=1}^{n} D(\mathbf{p}_i, r_i)$. We can rewrite the objective of Equation 6 in

following recursive form:

$$
\begin{aligned}
d\left(\bigcup_{i=1}^{n} \bar{D}_I(\mathbf{p}_i, r_i), I\right) &= d\left(\bar{D}_I^{n-1} \cup \bar{D}_I(\mathbf{p}_n, r_n), I\right) \\
&= d(\bar{D}_I^{n-1}, D_I^{n-1}) + d(\bar{D}_I(\mathbf{p}_n, r_n), D_I(\mathbf{p}_n, r_n)) \\
&\quad - d(\bar{D}_I^{n-1} \cap \bar{D}_I(\mathbf{p}_n, r_n), D_I^{n-1} \cap D_I(\mathbf{p}_n, r_n))) \qquad (8)
\end{aligned}
$$

5 Notes, Q&A

[*stavros*: can we prove that this leads to an optimal solution?]
PROBABLY YES! Most pixels in the image will be covered (explained) by multiple, neighbouring disks. *HOWEVER*, the pixels that are covered *only* by one disk are the ones that help us build the optimal substructure.

[*stavros*: Is DP necessary? Can we use a greedy approach?]
For the ideal case where we have only uniform neighbourhoods of discrete RGB values, and an additive distance function, it seems we can use a greedy approach. But what happens in the case of overlapping disks? [*update*: Also look at the *geometric set cover* problem.]

[*stavros*: Is is the same thing trying to minimize the reconstruction error and trying to find discontinuities in the encodings of concentric disks to determine medial points?]
Using a "simple" generator function g the reconstructs the disk neighbourhood around it by replicating a single RGB value, can identify points that correspond to maximal disks, in a similar way to identifying discontinuities in the encodings f of concentric disks.

[*stavros*: Should I enforce the generator function to reconstruct the disk uniformely, using a single RGB value?]
It is not necessary but if we want our problem to be well-defined, we may have to aim for local homogeneity, i.e. consider that the input image can be viewed as the union of locally homogeneous regions. In that case, the generator function should be as "simple" as possible, e.g. replicating a single RGB value across the whole disk neighbourdhood.

[*stavros*: How to combine the reconstructions that stem from neighbouring, overlapping disks?]
This proves to be one of the main problems with our formulation. For now we use the simple approach of averaging the reconstructions (sum over the number of times a pixel has been covered by some disk.).

[*stavros*: What happens if I want to detect axes for a single scale (radius)?]
We have two options depending on whether we want to be able to fully reconstruct the image, or just detect medial points that *precisely* correspond only to the wanted scale. In the former case, we will have very "thick" medial axes, that are at most r pixels away from a boundary. In the latter case, our formulation does not work. We have to compare disk responses at neighbouring scales $r+1$ and $r-1$ and find discontinuities. However, we could alternatively use the first approach, and recover the medial points that correspond to scale r by seeking axes up to a certain thickness, close to the scale of interest.

[*stavros*: What happens if the generator function g has infinite capacity?]
If g has infinite capacity, that means that, regardless how complex its neighbourhood is (highly textured, noise, complex illumination and occlusion effects), it can reconstruct it perfectly. In that case, our formulation is not useful, as we would be able to perfectly reconstruct even disks centred on boundary points, which is not compatible with the notion of a MAT.

[*stavros*: How to model the notion of the maximal inscribed disk for color images?]
We can apply an initial pruning step, at which we compute f for every point in the scale-space and then $\forall D(\mathbf{p}_i, r_i)$ check if there is a disk $D(\mathbf{p}_j, r_j)$ with $r_j > r_i$ and with very similar decodings: $d(\bar{D}_I(\mathbf{p}_i, r_i), \bar{D}_I(\mathbf{p}_j, r_j)) < \epsilon$. In that case, we remove $(\mathbf{p}_i, r_i)$ from the set of candidate medial points.

The idea is that small disks in the interior of the object will not be selected because they will be subsumed by other, larger (maximal) circles in the interior of the object. Circles that cross object boundaries will not be selected because the will not be able to accurately reconstruct the local patch well enough (except if g has very high capacity - see above).