Computers, Environment and Urban Systems 57 (2016) 48–58

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Characterizing the shapes of noisy, non-uniform, and disconnected point clusters in the plane

Xu Zhong [a,*], Matt Duckham [b]

[a] Department of Infrastructure Engineering, University of Melbourne, Victoria 3010, Australia
[b] School of Mathematical and Geospatial Sciences, RMIT University, Victoria 3001, Australia

ARTICLE INFO

Article history:
Received 6 December 2015
Received in revised form 12 January 2016
Accepted 16 January 2016
Available online 3 February 2016

Keywords:
Footprint
Non-convex
Delaunay triangulation
Clustering
α-Shape
χ-Shape
DBSCAN

ABSTRACT

Many spatial analyses involve constructing possibly non-convex polygons, also called "footprints," that characterize the shape of a set of points in the plane. In cases where the point set contains pronounced clusters and outliers, footprints consisting of disconnected shapes and excluding outliers are desirable. This paper develops and tests a new algorithm for generating such possibly disconnected shapes from clustered points with outliers. The algorithm is called χ-outline, and is based on an extension of the established χ-shape algorithm. The χ-outline algorithm is simple, flexible, and as efficient as the most widely used alternatives, $O(n\log n)$ time complexity. Compared with other footprint algorithms, the χ-outline algorithm requires fewer parameters than two-step clustering-footprint generation and is not limited to simple connected polygons, a limitation of χ-shapes. Further, experimental comparison with leading alternatives demonstrates that χ-outlines match or exceed the accuracy of α-shapes or two-step clustering-footprint generation, and is more robust to some forms of non-uniform point densities. The effectiveness of the algorithm is demonstrated through the case study of recovering the complex and disconnected boundary of a wildfire from crowdsourced wildfire reports.

1. Introduction

Many real-life applications require the construction of polygonal regions that characterize the distribution of a set of points in the plane $P \subset \mathbb{R}^2$ (e.g., Downs and Horner (2009)). Such regions are called "footprints" of *P*. A typical structure used to generate a footprint is the *convex hull*. The convex hull of *P* is the smallest convex polygon that contains all points in *P* (De Berg, Van Kreveld, Overmars, & Schwarzkopf, 2000). However, in cases where the distribution of points is markedly non-convex, there can be no single "correct" footprint. Rather, the accuracy of a footprint may depend on the specific application, or on human cognition and preference. Further, the point distribution may be best characterized by disconnected polygons, possibly neglecting outliers (Lee & Estivill-Castro, 2006; Lee, Qu, & Lee, 2012). Fig. 1 shows an example of a point set *P* that contains pronounced clusters and outliers. In such a case the convex hull significantly fails to capture the shape of *P*, illustrated in Fig. 1(a).

Today, several algorithms exist to construct "accurate" footprints for such non-convex and clustered point distributions. Because there can be no unique non-convex polygon, such algorithms require at least one adjustable parameter to obtain desirable footprints. This paper develops and tests a new algorithm for generating possibly disconnected polygons that characterize the shapes of such non-convex and clustered point distributions. The algorithm, called χ-outline is an adaptation of the established χ-shape algorithm (Duckham, Kulik, Worboys, & Galton, 2008) to handle disconnected shapes and outliers using only a single parameter. Fig. 1(b) illustrates a typical output of the algorithm.

Following a review of the background literature in Sections 2 and 3 describes the χ-outline algorithm itself in detail. Section 4 then evaluates the performance of the algorithm against the two leading alternatives: α-shapes, and a two-step clustering-footprint generation based on DBSCAN and χ-shapes. The evaluation shows that in most cases the accuracy of χ-outlines equals or outperforms these alternatives, in particular where individual clusters tend to differ in point densities. Section 5 illustrates the application of χ-outlines to a case study of wildfire perimeter estimation based on crowdsourced fire reports, where systematic differences in cluster densities are common, due to variations in population density. Finally, Sections 6 and 7 provide a discussion of the results and the final conclusions, respectively.

2. Background

This paper focuses on footprint algorithms where the goal is to construct a polygonal region that adequately represents the distribution of a given set of points, *P*, that occupies one or more regions in the plane. We do not here consider the special case where points in *P* lie only on curves or the boundary of regions. Numerous algorithms already exist to generate curves or outlines for this latter case (e.g. Amenta, Choi, & Kolluri, 2001a, 2001b, Attali, 1997, Traka & Tziritas, 2003).

* Corresponding author.
E-mail address: zhong.xu@unimelb.edu.au (X. Zhong).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.01.003
0198-9715/

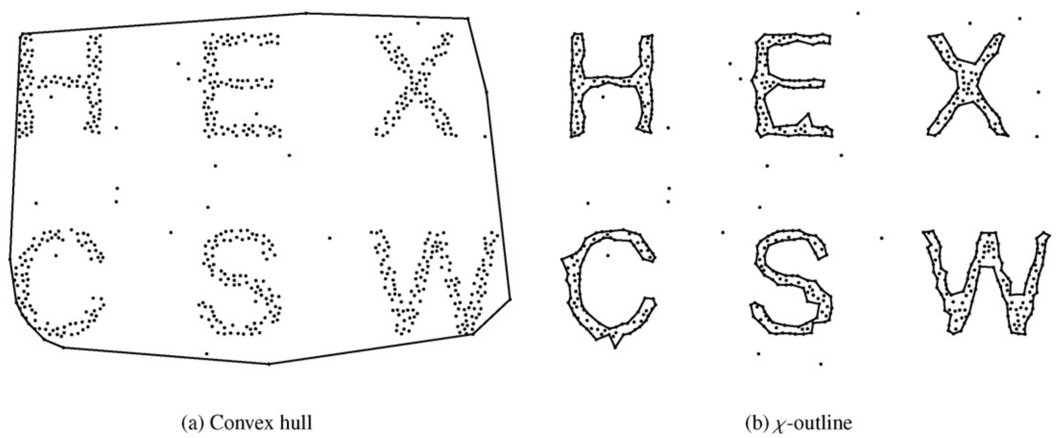

Fig. 1. (a) Convex hull and (b) χ-outline ($\chi = 0.33$) of an example set of two-dimensional points. The points contain pronounced clusters, outliers, and illustrate a non-convex distribution.

2.1. Simple region-based footprints

One characteristic frequently used to distinguish footprint algorithms is whether or not all the input points in P are required to lie in the interior of the footprint. For example, *covering discs* and the *covering discs with tangents* methods (Galton & Duckham, 2006) assign an influence region to each point in P. The footprint is the union of the influenced regions of all points in P. The shape and size of the influenced region decide the resulting footprint, but all input points in P will be contained in the footprint.

The Voronoi diagram (VD) based method of Alani, Jones, & Tudhope (2001) and the Delaunay triangulation (DT) based method of Arampatzis et al. (2006) also produce footprints that contain all the input points. These methods additionally require auxiliary points P_α which are considered to be *outside* of the footprint. The principle is to construct a footprint whose edges separate a point in P from its neighboring points in P_α, so that the footprint contains all points in P and excludes all points in P_α. The edges of the VD and the DT of $P \cup P_\alpha$ are used to generate the edges of the footprint. In the VD based method, Voronoi edges that are shared by a Voronoi cell of a point in P and a Voronoi cell of a point in P_α form the footprint (Alani et al., 2001). In the DT based method, the edges of the footprint are generated by connecting the midpoints of the triangulation edges that connect a point in P and a point in P_α (Arampatzis et al., 2006). These two methods are most appropriate in cases where P_α is given, i.e., where the problem involves a set of both positive (included) and negative (excluded) points. Galton and Duckham (2006) described an iterative approach to using these two methods when P_α is not given. An initial P_α is generated randomly outside the convex hull of P to produce a preliminary footprint. Then a new P_α is generated outside the preliminary footprint, which produces a new footprint. The final footprint can be obtained by a pre-defined number of successive iterations.

In many cases, footprint algorithms generate a single, simple polygon to characterize the shape of P. For example, the k-nearest neighbors (kNN) based method (Moreira & Santos, 2007) was proposed by generalizing the gift-wrapping convex hull algorithm (Jarvis, 1973). At each iteration, the kNN-based algorithm finds the next point from the kNN of current points, instead of processing the entirety of P used in the gift-wrapping algorithm. Each subsequent point produces the largest right-hand turn from the current point without resulting in self-intersection. When no legal subsequent points exist, the algorithm has to increase k and rerun from the beginning. The algorithm also increases k and reruns from the beginning if it generates a footprint that does not include all points in P. Hence the algorithm has a poor worst-case efficiency. The χ-shape algorithm (introduced in detail in Section 3) proposed by Duckham et al. (2008) is more efficient than the kNN-based method, but also can only produce a single, simple polygonal footprint.

2.2. Pre-clustering and outliers

To characterize more complicated and clustered point distributions, such as illustrated in Fig. 1, input points may be pre-processed using a spatial clustering algorithm (e.g. Miller & Han, 2009, Xu & Wunsch, 2005). In general, clustering may be able to handle a priori unknown numbers of clusters with arbitrary shapes, as well as outliers. For example, one of the most widely used spatial clustering algorithms is the density-based DBSCAN (Ester, Kriegel, Sander, & Xu, 1996). Clustering using DBSCAN can partition the point set P into one or more disjoint clusters P_i and possibly a set of outliers O. After clustering, any of the footprint algorithms discussed above, including the kNN-based algorithm or the χ-shape algorithm, can be applied to each cluster independently.

The approach of pre-clustering process allows independent identification of clusters and outliers, providing great flexibility in footprint generation. But there are two disadvantages. First, any pre-clustering algorithm necessarily requires additional parameterization. Just as for non-convex footprint generation, there can be no single correct answer for clustering. Selecting the correct clustering parameter may be difficult to achieve automatically. The quality of the constructed footprints is then strongly dependent on the parameterization of the clustering algorithm. Second, depending on the footprint algorithm parameterization, the footprints of clusters may intersect with one other. In this case, the union of the footprints needs to be calculated to obtain regular polygonal shapes. In addition, the union of the footprints possibly contains holes, which also need to be detected and removed for regular polygonal shapes. These additional steps may also increase the complexity and computational overhead of this approach.

2.3. Footprints for clusters and outliers

Some well-known footprint algorithms do enable the construction directly of disconnected polygonal shapes for points with clustered distributions. The output footprints of these algorithms may consist of polygons, lines, and isolated points. The polygonal shapes may not contain all points in P. These algorithms can separate sampling points (points contained in the polygonal shapes) from outliers (points not contained in the polygonal shapes) by themselves, without the need of pre-clustering. The α-shape algorithm (Edelsbrunner, Kirkpatrick, & Seidel, 1983) is the most famous example, which generalizes the convex hull with a single parameter α, the multiplicative inverse of the radius of closed disks. Negative parameters yield the complement of an open disk of radius $-1/\alpha$. The α-shape of P is a sub-graph of the DT of P. For large negative α, the α-shape of P is just P itself. When $\alpha = 0$, the α-shape is the convex hull of P.