COMPLEX CURVE TRACING BASED ON A MINIMUM SPANNING TREE MODEL AND REGULARIZED FUZZY CLUSTERING

Benson S. Y. Lam[1] and Hong Yan[1,2]

[1]Department of Computer Engineering and Information Technology
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[2]School of Electrical and Information Engineering
University of Sydney, NSW 2006, Australia

ABSTRACT

The fuzzy curve-tracing (FCT) algorithm can be used to extract a smooth curve from unordered noisy data. However, the model produces good results only if the curve shape is either opened or closed. In this paper, we propose several techniques to generalize the FCT algorithm for tracing complicated curves. We develop a modified clustering algorithm that can produce cluster centers less dependent on the pre-specified number of clusters, which makes the reordering of cluster centers easier. We make use of the Eikonal equation and the Prim's algorithm to form the initial curve, which may contain sharp corners and intersections. We also introduce a more powerful curve smoothing method. Our generalized FCT algorithm is able to trace a wide range of complicated curves, such as handwritten Chinese characters.

1. INTRODUCTION

Extracting a smooth curve from a data set has found many applications to handwriting recognition, line pattern features extraction and skeletonization. Due to the presence of noise and variations in shape, curve tracing is often a complicated problem. A number of algorithms have been developed to solve this problem and they can be grouped into the following three classes.

The first class includes principal curve [1] and fuzzy curve tracing [2] based methods. In these methods, a mathematical model and user-defined initial guess are necessary. These models contain a regularization term and are robust against noise. As the user-defined curve is needed, prior knowledge about the object is necessary.

The second class includes data-based methods, such as medial axis transform, Voronoi diagram and their variants [3]. These methods do not use mathematical models. An advantage of these methods is that the shape of the object can be an unknown and they can deal with complicated curve shapes. However, these approaches are often sensitive to noise.

The third class is based on hybrid approaches, which combine both curve-based and data-based methods. Usually, a data-based method is applied first to get a curve. Then, this curve is used as an initial curve in the curve-based algorithm. The method in [4] belongs to this class. However, this method requires many user-defined thresholds.

In this paper, we developed several techniques to generalize the FCT algorithm for tracing complicated curves, which combines advantages of curve and data based techniques. The proposed method consists of three steps. Firstly, a modified fuzzy c-means (FCM) algorithm is introduced, which partitions the data set into groups. Then, a cluster center reordering process is carried out to formulate the centers as a curve. The main idea of the method is to form a minimal spanning tree (MST) from the data, based on the Eikonal equation [5] and Prim's algorithm [6]. The MST can produce complicated initial curve that can contain sharp corners and intersections. Finally, the curve is smoothed based on regularized fuzzy clustering.

The proposed method is able to extract a smooth complicated curve in a binary image. The algorithm cannot deal with gray-scale or color images directly, but it is applicable if input image is binarized using edge detection and thresholding or segmentation to extract curve-shaped regions. It is especially useful for processing of document images, which are usually binary. For example, it can be used for detection of curved text, thinning or skeletonization of line drawings and off-line recognition of handwritings. Although the effectiveness of the proposed algorithm is demonstrated with two-dimensional images in this paper, the method can be easily generalized to deal with higher dimensional data.

2. DATA PARTITION USING A MODIFIED FCM ALGORITHM

In this section, we introduce a modified FCM algorithm, which partition the data into groups.

Assume the input data samples are $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$. The data samples are divided into c classes and each class is represented by a cluster center $\mathbf{v}_k \in V = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ where $1 \le k \le c$. The modified FCM algorithm is given as

0-7803-8554-3/04/$20.00 ©2004 IEEE.

$$J_1(\mathbf{U},\mathbf{V};\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{m}\left|\mathbf{x}_i - \mathbf{v}_k\right|^2 + \alpha\sum_{\substack{k=1\\(k\neq q)}}^{c}\sum_{q=1}^{c}\omega_{kq}^{m}\left|\mathbf{v}_k - \mathbf{v}_q\right|^2 . \quad (1)$$

where m is a constant and taken as 2 in this paper, α is a threshold and ω_{kq} is a weighting function satisfying the possibilistic operator in [7]. The modified FCM algorithm consists of two terms. The first term is from FCM algorithm and thus will drive the cluster centers towards the mean of the data. The second term is a product of $|\mathbf{v}_k - \mathbf{v}_q|^2$ and a weighting function ω_{kq}. The weighting function ω_{kq} can be written as

$$\omega_{kq} = \frac{1}{1+\alpha\left|\mathbf{v}_k - \mathbf{v}_q\right|^2} . \quad (2)$$

This product term in Equation (1) provides a simple strategy adjusting the total number of cluster centers. ω_{kq} is of the form $1/(1+x^2)$. The function $1/(1+x^2)$ is closed to 1 if x is closed to zero. If $|x|$ is large, $1/(1+x^2)$ will be nearly zero. Therefore, the new term will be valid only if $\mathbf{v}_k$ and $\mathbf{v}_q$ are closed to each other. If the total number of cluster centers is large in the data set, there will be many cluster centers produced. Based on the property of the modified FCM algorithm, two close clusters will be dragged towards to each other. Therefore, we can adjust the total number of cluster centers n by the following scheme.

If $|\mathbf{v}_k - \mathbf{v}_q| \le \varepsilon$, then delete $\mathbf{v}_q$ and decrease n by 1.

3. RELATIONAL GRAPH USING DISTANCE MATRIX BASED ON EIKONAL EQUATION

In this section, we introduce a relational graph, which connects the cluster centers obtained from Section 2. The key idea of this method is to connect the cluster centers using minimal spanning tree (MST) based on the Eikonal equation [5] and Prim's algorithm [6].

We adopt the Eikonal equation to obtain the MST. The Eikonal equation is constructed as follows. We first define a surface T(x,y) on domain Ω as

$$T(x_i,y_i) = \begin{cases} 1 - \max_k \mu_k(x_i,y_i) & \text{if } (x_i,y_i)\in X \\ \theta_a & \text{otherwise} \end{cases}, \quad (3)$$

where θ_a is a threshold taken as 5 in our experiments, the domain Ω is taken as the smallest rectangular embedding the data X and $\mu_k(x_i,y_i)$ is the membership value of the data (x_i,y_i). The shortest path problem is formulated as

$$u(x,y) = \min_{\gamma}\int_{\mathbf{v}_k}^{(x,y)} T(\gamma(\tau))d\tau , \quad (4)$$

with boundary condition $u(v^x_k, v^y_k) = 0$ and $\mathbf{v}_k=(v^x_k, v^y_k)$. The variable u represents the shortest distance from point (x_i,y_i) to $\mathbf{v}_k$. If we consider the difference between any two very close points on u(x,y), Equation (4) can be formulated as

$$|\nabla u(x,y)| = T(x,y). \quad (5)$$

This is the standard form of the Eikonal equation. In this equation, each data $T(x_i,y_i)$ can be viewed as the total distance needed for traveling through the data point (x_i,y_i). If a data point (x_s,y_s) is close to the cluster centers $\mathbf{v}_k$, $u(x_s,y_s)$ will be small. Conversely, if a data point (x_f,y_f) is far away or even does not lie in the set X, $u(x_f,y_f)$ will be large. Figure 1(b) shows the surface $T(x_i,y_i)$ for the data in Figure 1(a). In this figure, the gray scale of each pixel represents the distance. The dark regions represent small traveling distances through the corresponding point while the light regions represent large traveling distances through the corresponding point.

Now, the shortest distance between two nodes $\mathbf{v}_p$ and $\mathbf{v}_q$ on Ω can be evaluated based on algorithm 1.

Algorithm 1:

Initialization: Set i=1 and c=$c_0$where c_0 is a threshold and taken to be 0.1 in our experiments.

Let $F(x_i,y_i,\mathbf{v}_p) = \int_{\mathbf{v}_p}^{(x_i,y_i)} T(x,y)dxdy$, where $(x_i,y_i)\in\Omega$.

Step 1: If every pair of nodes has been assigned a distance value, stop.
Otherwise, calculate $G(\mathbf{v}_p,c) = \{(x_i,y_i) \mid F(x_i,y_i,\mathbf{v}_p) \le c\}$.

Step 2: If $G(\mathbf{v}_p,c) \cap \mathbf{v}_q \neq \phi$, then the distance between $\mathbf{v}_p$ and $\mathbf{v}_q$ is c.
Otherwise, i=i+1, c=i*c_0 and go to step 1.

In Algorithm 1, the functions can be interpreted as follows. Since the function $T(x_i,y_i)$ can be viewed as the traveling distance through point (x_i,y_i). The function $F(x_i,y_i,\mathbf{v})$ can be interpreted as the traveling distance from node $\mathbf{v}_p$ to (x_i,y_i). The function $G(\mathbf{v}_p,c)$ is the set containing the nodes from which the traveling distance to $\mathbf{v}_p$ is less than c. Therefore, the algorithm models the shortest path problem from every pair of nodes and introduces a searching criterion for node $\mathbf{v}_p \in V$ for finding its nearest nodes with step size c_0.

After obtaining the distance matrix, two more procedures have to be taken. The first is to find the MST. This can be achieved using Prim's algorithm. Figure 1(c) shows the MST from the data in Figure 1(a). The second one is the introduction of an extra edge. MST can only contain n-1 edges. For general curve shape, we introduce the following edge insertion scheme. Let Ω_{insert} be the collections of all cluster centers, which contain only 1 edge. For each cluster center with $\mathbf{v}_p\in\Omega_{insert}$,

If $F(v^x_p, v^y_p, \mathbf{v}_q)<\theta$, $\mathbf{v}_p$ and $\mathbf{v}_q$ are connected.

where θ is a user-defined threshold, $\mathbf{v}_p=(v^x_p, v^y_p)$ and $\mathbf{v}_q$ is the cluster center which does not connect to $\mathbf{v}_p$.

4. CURVE SMOOTHNESS CONSTRAINT

In this section, we conduct a smoothing procedure to the curve $\mathbf{u}_k$ obtained from Section 3.

Since we study complicated curve shape, the smoothing procedure is formulated as

$$J_2 = \sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{m}|\mathbf{x}_i - \mathbf{v}_k|^2 + \alpha_1\sum_{k=1}^{c}\sum_{(p,q)\in\Gamma_k}\eta_{pqr}^{m}|\mathbf{v}_p - 2\mathbf{v}_k + \mathbf{v}_q|^2 + \beta\sum_{k=1}^{c}|\mathbf{v}_k - \mathbf{u}_k|^2 \quad (6)$$

where α_1 and β are weighting constants, Γ_k is the set of pairs $(\mathbf{v}_p,\mathbf{v}_q)$ of all cluster centers connected to $\mathbf{u}_k$. If there is only one node connected to $\mathbf{u}_k$, this set is empty. η_{pkq} satisfies the following constraints.

$$0 < \eta_{pkq} < 1 \qquad 0 \le k \le c \text{ and } (p,q) \in \Gamma_k \quad (7)$$

$$\sum_{(p,q)\in\Gamma_k}\eta_{pkq} = 1 \qquad 0 \le k \le c \quad (8).$$

The model in Equation (6) consists of three terms. The first term is the cost function in the FCM algorithm while the second term is the smoothness constraint. Note that a fuzzy operator η_{pkq} is imposed in the smoothness constraint. This can be interpreted as follows. $\mathbf{v}_p$ and $\mathbf{v}_q$ are any two neighbourhoods of the node $\mathbf{v}_k$. If the term $|\mathbf{v}_p-2\mathbf{v}_k+\mathbf{v}_q|^2$ is small compared with other pairs in Γ_k, η_{pkq} will become large and vice versa. Thus, smoothing will be conducted heavily on the one having the smallest second order difference. This can avoid smoothing out important features. For example, in Figure 1(c), "+" structure is available. Based on the above interpretations, smoothing of "+" will be conducted heavily to "— " and "|". The third term measures the degree of similarity between the solution curve and the original curve $\mathbf{u}_k$. This term can also preserve the structure of the curve.

Now, we apply this smoothing scheme to the data in Figure 1(a) and the result is given by Figure 1(d). We can see that the "+" structure can be preserved and a smooth curve is obtained without distortions.

5. EXPERIMENTAL RESULTS

In this section, we further test the robustness of the method in three experiments.

The first experiment is devoted to the curved text image in Figure 1(e) and it is corrupted by binary noise. After applying the proposed method, a smooth curve is extracted, which is given by Figure 1(f).

In the second experiment, we consider an open square shape in Figure 1(g). The curve tracing result using FCT algorithm is given by Figure 1(h). Since the gap on the square is too small, the FCT algorithm misclassified it as an edge. Figure 1(i) shows the result using the proposed method. Since the curve is connected based on shortest path, the gap is retained and an accurate result is yielded.

Results from the third experiment are shown in Figure 1(j). This data set represents a handwritten Chinese character with binary noise corruption. The curve tracing result using the proposed method is given in Figure 1(k). Although the data is corrupted by noise and the shape is quite complicated, the proposed method is able to produce an accurate result.

6. CONCLUSION

In this paper, we have introduced a method for tracing complex curves based on an MST model and regularized fuzzy clustering. Our method is insensitive to the specified number of clusters and can reorder cluster centers to form meaningful curves. The proposed algorithm can resolve line cross sections and does not produce any artifacts. The method generates accurate results even noise and broken structures are present.

The computational complexity of our method is analyzed as follows. In Algorithm 1, the Eikonal equation is used to find the distance matrix. It has *n* steps, where *n* is the number of clusters. In each step, we have to take the summation of all the elements in the path from $\mathbf{v}_p$ to (x_i,y_i). Since only the summation is taken into account, the computational complexity is not high. The complexity of Prim's algorithm is $O(n^2)$. The time needed for data clustering and curve smoothing is similar to that of the original FCT algorithm.

Acknowledgement: This work is supported by Hong Kong Research Grant Council (project CityU 1035/02E).

7. REFERENCES

[1] Trevor Hastie and Werner Stuetzle, "Principal Curves", *Journal of the American Statistical Association*, Vol. 84, No. 406, Jun., 1989

[2] H. Yan, "Fuzzy curve-tracing algorithm." *IEEE Transactions. On systems, MAN, and Cybernetics*, Vol. 31, no. 5, pp. 768-780, 2001.

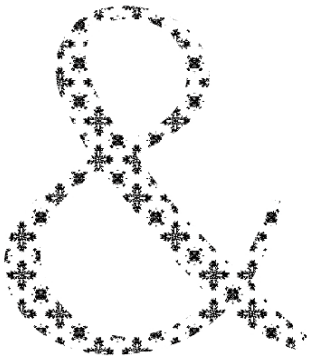
[3] Paul Yushkevich, P. Thomas Fletcher, Sarang Joshi, Andrew Thall and Stephen M. Pizer, "Continuous medial representations for geometric object modeling in 2D and 3D", *Image and Vision Computing*, Vol 21, Issue 1, pp. 17-27, Jan 2003.

[4] Kegl, B. and Krzyzak, A, "Piecewise linear skeletonization using principal curves", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 24, Issue 1, pp. 59-74, Jan. 2003.

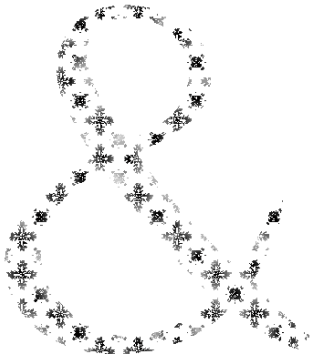
[5] J. A. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision and Materials Sciences*, Cambridge: Cambridge University Press, 1996.

[6] R. Gounld, *Graph Theory*, Menlo Park, Calif.: Benjamin/Cummings, 1988.

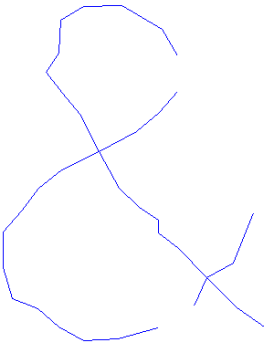
[7] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering." *IEEE Transactions on Fuzzy Systems*, Vol. 1, no. 2, pp. 98-110, 1993.



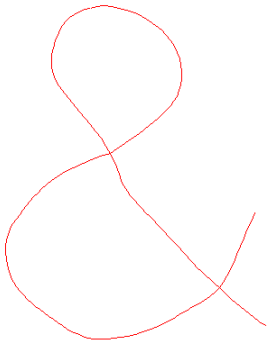
(a) “&” symbol.



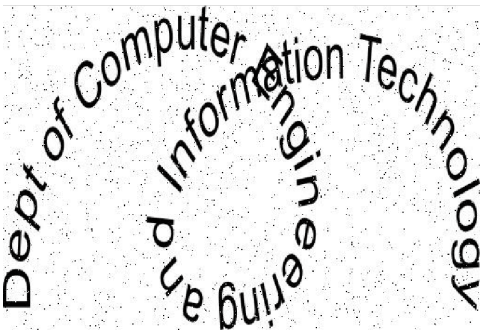
(b) The surface $T(x_i,y_i)$.



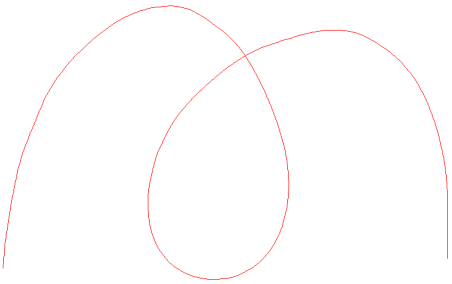
(c) MST of (a) using relational graph.



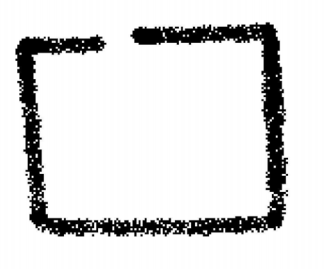
(d) Curve extracted using the proposed method.



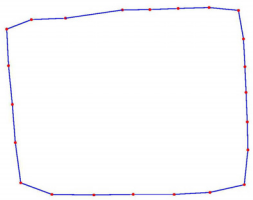
(e) Curved text image.



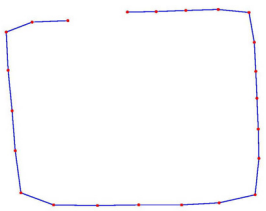
(f) Curve extracted using the proposed method.



(g) An open Square data.

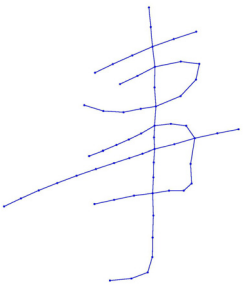


(h) Curve extracted using FCT algorithm.



(i) Result using the proposed method.

(j) A Chinese character with binary noise.



(k) Result using the proposed method.

Figure 1. Curve extraction using the proposed method.