Computational Geometry 47 (2014) 625–641

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

Similarity of polygonal curves in the presence of outliers

Jean-Lou De Carufel [a,1], Amin Gheibi [a,2,3,*], Anil Maheshwari [a,3], Jörg-Rüdiger Sack [a,2,3], Christian Scheffer [b,4]

[a] *School of Computer Science, Carleton University, Ottawa, ON, Canada*
[b] *Department of Computer Science, Westfälische Wilhelms-Universität Münster, Germany*

ARTICLE INFO

Article history:
Received 23 March 2013
Received in revised form 16 September 2013
Accepted 3 January 2014
Available online 8 January 2014
Communicated by R. Klein

Keywords:
Fréchet distance
Similarity of polygonal curves
Approximation
Weighted shortest path

ABSTRACT

The Fréchet distance is a well studied and commonly used measure to capture the similarity of polygonal curves. Unfortunately, it exhibits a high sensitivity to the presence of outliers. Since the presence of outliers is a frequently occurring phenomenon in practice, a robust variant of Fréchet distance is required which absorbs outliers. We study such a variant here. In this modified variant, our objective is to minimize the length of subcurves of two polygonal curves that need to be ignored (MinEx problem), or alternately, maximize the length of subcurves that are preserved (MaxIn problem), to achieve a given Fréchet distance. An exact solution to one problem would imply an exact solution to the other problem. However, we show that these problems are not solvable by radicals over $\mathbb{Q}$ and that the degree of the polynomial equations involved is unbounded in general. This motivates the search for approximate solutions. We present an algorithm which approximates, for a given input parameter δ, optimal solutions for the MinEx and MaxIn problems up to an additive approximation error δ times the length of the input curves. The resulting running time is $O(\frac{n^3}{\delta}\log(\frac{n}{\delta}))$, where n is the complexity of the input polygonal curves.

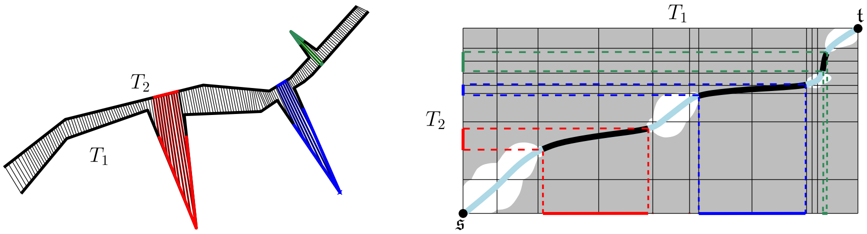
1. Introduction

Measuring similarity between two polygonal curves in the Euclidean space is a well studied problem in computational geometry, both in practical and theoretical settings. It is of practical relevance in areas such as pattern analysis, shape matching and clustering. It is of theoretical interest as well since problems in this domain tend to be challenging and often lead to innovative tools and techniques. The Fréchet distance – one of the widely used measures for similarity between curves – is intuitive and takes into account global features of the curves instead of local ones, such as their vertices [2,4,10]. Despite being a high quality similarity measure for polygonal curves, it is very sensitive to the presence of outliers. Consequently, research has been carried out to formalize the notion of similarity among a set of polygonal curves that tolerate outliers. It is based on intersection of curves in local neighborhood [17], topological features [5], or adding flexibility to incorporate the existence of outliers [10]. In [10], Driemel and Har-Peled discuss a new notion of robust Fréchet distance, where they allow up to k shortcuts between vertices of one of the two curves, where k is a constant specified as an input

* Corresponding author.
E-mail addresses: jdecaruf@cg.scs.carleton.ca (J.-L. De Carufel), agheibi@scs.carleton.ca (A. Gheibi), anil@scs.carleton.ca (A. Maheshwari), sack@scs.carleton.ca (J.-R. Sack), Christian.Scheffer@uni-muenster.de (C. Scheffer).

[1] Research supported by Fonds de Recherche du Québec – Nature et Technologies.
[2] Research supported by High Performance Computing Virtual Laboratory and SUN Microsystems of Canada.
[3] Research supported by Natural Sciences and Engineering Research Council of Canada.
[4] A part of this research was carried out while the author was visiting Carleton University.

0925-7721/$ – see front matter
http://dx.doi.org/10.1016/j.comgeo.2014.01.002



(a) Two polygonal curves T_1 and T_2. (b) Free-space diagram for T_1 and T_2.

Fig. 1. (a) A possible solution is illustrated by the connecting lines between the parameterizations for T_1 and T_2. The subcurves on both polygonal curves that should be ignored are illustrated by the blue, red and green subcurves on T_1 and T_2. So, $Q^B(T_1, T_2)$ is the summation of the lengths of the colored subcurves and $Q^W(T_1, T_2)$ is that of the black subcurves. (b) The solution corresponds to an xy-monotone path in the deformed free-space diagram F. In this space, $Q^B(T_1, T_2)$ can be measured by summing the lengths of its subpaths going through the forbidden space (shaded gray area), measured in the L_1-metric (similarly for $Q^W(T_1, T_2)$) – see Section 1.2 for definitions of $Q^B(T_1, T_2)$ and $Q^W(T_1, T_2)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parameter. They provide a constant factor approximation algorithm for finding the minimum Fréchet distance among all possible k-shortcuts. One drawback of their approach is that a shortcut is selected without considering the length of the ignored part. Consequently, such shortcuts may remove a significant portion of a curve. As a result, substantial information about the similarity of the original curves could be ignored. A second drawback of their approach is that the shortcuts are only allowed on one of the curves. Since noise could be present in both curves, shortcuts may be required on both to achieve a good result. For example, Fig. 1(a) shows two polygonal curves that both need simultaneously shortcuts to become similar.

In this paper, we discuss an alternative Fréchet distance measure that tolerates outliers. It incorporates the lengths of the curves and allows the possibility of shortcuts on both curves. We consider two natural dual perspectives of this problem. They are outlined as follows using the common dog-leash metaphor for Fréchet distance.

Assume that a person wants to walk along one curve and his/her dog on another one. Let $\varepsilon \geqslant 0$ be a given constant. The *Min-Exclusion* (MinEx) *Problem* is to determine a walk that minimizes the total length of all parts of the curves for which a leash of length bigger than ε is needed. The *Max-Inclusion* (MaxIn) *Problem* is to determine a walk that maximizes the total length of all parts of the curves for which a leash of length at most ε is sufficient. Observe that the solution for the MinEx problem leads to a solution for the MaxIn problem and vice-versa. An exact solution for these problems is presented in [6], where the distances are measured using (more restrictive and simpler) L_1 and L_∞ metrics. In particular, they solved exactly the MinEx (respectively the MaxIn) problem under L_1 and L_∞ metrics in $O(n_1 \cdot n_2 \cdot (n_1 + n_2) \cdot \log(n_1 \cdot n_2))$ time, using $O(n_1 \cdot n_2 \cdot (n_1 + n_2))$ space, where n_1 and n_2 denote the complexity of the input polygonal curves T_1 and T_2, respectively.

In Section 2, using Galois theory, we show that these problems are not solvable by radicals over $\mathbb{Q}$, when distances are measured using the L_2-metric. (It is natural to study Fréchet distance problems in L_2-metric, see e.g. [2].) This suggests that we should look for approximation algorithms. Har-Peled and Wang [12] proposed an approximation algorithm for the MaxIn problem whose running time is $O(\frac{n^4}{\delta^2})$, where n is the size of the input polygonal curves. They claimed that their approach is a $(1-\delta)$-approximation, that is, the accuracy of their solution is greater than $(1-\delta)$ times the optimal one [12, Theorem 4.2]. However, we show via a counterexample that this claim is not true (see Appendix A and [18]). Therefore, to the best of our knowledge, no FPTAS (Fully Polynomial-Time Approximation Scheme) exists for this problem. In this paper, we provide algorithms that approximate solutions for the MinEx and MaxIn problems up to an additive approximation error δ times the length of the input curves.

1.1. Preliminaries

Let $T_1 : [0, 1] \to \mathbb{R}^2$ and $T_2 : [0, 1] \to \mathbb{R}^2$ be two polygonal curves. Their *Fréchet distance* is defined as the minimum *leash length* required to walk only forwardly, in parallel on both T_1 and T_2, from the starting points to the ending points. Note that the two walks could have different variating speeds. More formally, two monotone parameterizations $\alpha_1, \alpha_2 : [0, 1] \to [0, 1]$ define, for each time $t \in [0, 1]$, a matching $(T_1(\alpha_1(t)), T_2(\alpha_2(t)))$ of one point on T_1 to exactly one point on T_2 and vice-versa. The needed leash length for the two parameterizations is defined as the maximum Euclidean distance between two matched points, over all times. Then, the Fréchet distance $\delta_F(T_1, T_2)$ is defined as the infimum of the required leash lengths over all possible pairs of monotone parameterizations [2]:

$$\delta_F(T_1, T_2) := \inf_{\alpha_1, \alpha_2 : [0,1] \to [0,1]} \max_{t \in [0,1]} \{ |T_1(\alpha_1(t)) T_2(\alpha_2(t))| \}, \tag{1}$$

where $|\cdot|$ denotes the Euclidean distance. For simplification, we say that all considered parameterizations are monotone. The corresponding *Fréchet distance decision problem* asks whether there exist two parameterizations for a given leash of length ε, realizing a Fréchet distance between T_1 and T_2 that is upper bounded by ε. In other words, it asks whether it is possible for someone to walk his/her dog with a given leash of length ε, such that the person and the dog stay on their own curves.