21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

# Compression of GPS Trajectories using Optimized Approximation

Minjie Chen[1], Mantao Xu[2], Pasi Fränti[1]
[1]*University of Eastern Finland, Finland;* [2]*Shanghai Dianji University, China*
*E-mail: {mchen,franti}@cs.joensuu.fi, xumt@sdju.edu.cn*

## Abstract

*A large number of GPS trajectories, which include users' spatial and temporal information, are collected by geo-positioning mobile phones in recent years. The massive volumes of trajectory data bring about heavy burdens for both network transmission and data storage. To overcome these difficulties, GPS trajectory compression algorithm (GTC) was proposed recently that optimizes both the data reduction by trajectory simplification and the coding procedure using the quantized data. In this paper, instead of using greedy solution in GTC algorithm, the approximation process is optimized jointly with the encoding step via dynamic programming. In addition, Bayes' theorem is applied to improve the robustness of probability estimation for encoded values. The proposed solution has the same time complexity with GTC algorithm in the decoding procedure and experimental results show that its bit-rate is around 80% comparing with GTC algorithm.*

## 1. Introduction

Location-acquisition technologies, such as geo-positioning mobile devices, enable users to obtain their locations and record travel experiences by a number of time-stamped trajectories. In the location-based web services, users can record, then upload, visualize and share those trajectories [1].

However, these trajectories often incur a large amount of redundant storage to the end-users as well as the mobile service providers. For example, if data is collected at 10 second intervals, a calculation in [2] shows that without any compression, 100 Mb of storage capacity is required to store the GPS trajectories of 400 users for a single day in server side. To overcome these difficulties, a number of compression algorithms have been presented not only considering the data reduction for visualization purpose but also investigating the encoding process for the storage use.

Due to the inherent characteristics in GPS trajectories, conventional error measure, e.g. the perpendicular Euclidean distance is not suitable for GPS trajectories as both spatial and temporal information should be considered. Therefore, the so-called *top-down time-ratio* (TD-TR) algorithm [3] was developed, where synchronous Euclidean distance was used instead of the perpendicular distance in the Douglas-Peucker algorithm [2]. Threshold-guided algorithm was also proposed via estimating the safe area of the next point using the position, speed and orientation information [4]. In [5], a multi-resolution simplification algorithm has also been designed with linear time complexity. Two error measures, called *local integral square synchronous Euclidean distance* (LSSD) and *integral square synchronous Euclidean distance* (ISSD) are used jointly, which can be calculated in $O(1)$ time. Semantic meanings of the GPS trajectories are also considered during the compression process in urban area in [6] whereas trajectory compression algorithm with network constraint has been developed in [7]. Performance evaluations are also made for several traditional trajectory simplification algorithms [8]. It should be mentioned that there is not one algorithm that always outperforms other compression approaches in all situations. However, these methods lack a rigorous analytical approach on the encoding procedures of the reduced trajectories. Namely, fixed bits are allocated after data reduction to store latitude, longitude and timestamp information.

On the other hand, when encoding techniques are used, a better compression ratio is achieved for the spatial trajectory data, which is appropriate for data storage. For example, quantization-based approach has been analytically investigated in the so-called vector map compression problem [9, 10]. In these algorithms, differential coordinates of adjacent data points are

used as the prediction errors. These residual vectors are then quantized and encoded using a variety of quantization strategies. For GPS trajectories, in [11], speed information is used in arithmetic coding using a fixed prediction model in 2-D space. In [12], data reduction and the quantized speed and direction changes are combined to seek an encoded trajectory using greedy approximation process, which achieves a state-of-the-art compression result.

In this paper, the optimized GPS trajectory compression algorithm (OGTC) is proposed. Approximation result with the minimum coding cost is selected for encoding using an optimization process via dynamic programming. In addition, Bayes' theorem is applied in order to improve the robustness of the probability estimation for the encoded variants.

## 2. Lossy compression of GPS Trajectory

### *2.1 Quantization process*

In this paper, maximum synchronous Euclidean distance (max SED) [3] is used as the error measure to evaluate the distortion between the original and compressed GPS trajectories. The error is measured by the maximum synchronous distance between original positions and its synchronized approximated positions.

In vector map compression, differential coordinates are used directly in the encoding process. However, in GPS trajectories, these differential coordinates will be inconsistent after the data reduction (approximation) process. Meanwhile, speed and direction changes are more robust variants even if an approximation is made with different reduction rate in different segments.

Fig.1 is an example of the proposed quantization process. Suppose $p_i'$ and $p_j'$ are the quantized position for point $p_i$ and $p_j$, the sub-segment $P_i^j$ can be approximated by line segment $\overline{p_i' p_j'}$ when the approximation error $\delta_{SED}(P_i^j, \overline{p_i' p_j'})$ is less than the given error tolerance $\varepsilon$. Here, we set the quantization error for point $p_i$ and $p_j$ as $\gamma\varepsilon$ at maximum, where $\gamma$ = 0.5 is a parameter.

After the differential coordinates are quantized in polar space, given time interval $t_j$ - $t_i$, the quantized level of speed from $p_i'$ to $p_j$ can be calculated as:

$$l_v(i,j) = \sqrt{2} \cdot \gamma \cdot \varepsilon / (t_j - t_i) \quad (1)$$

Therefore, given $d(p_i', p_j)$, which is the distance between $p_i'$ and $p_j$, the quantized speed is calculated as:

$$v^*(i,j) = [d(p_i', p_j) / l_v(i,j)] \cdot l_v(i,j) \quad (2)$$

Meanwhile, the direction change $\Delta\theta(i, j)$ has a value between $-\pi$ and $\pi$. Given the quantized speed $v^*(i, j)$, the quantization level for the direction change can be

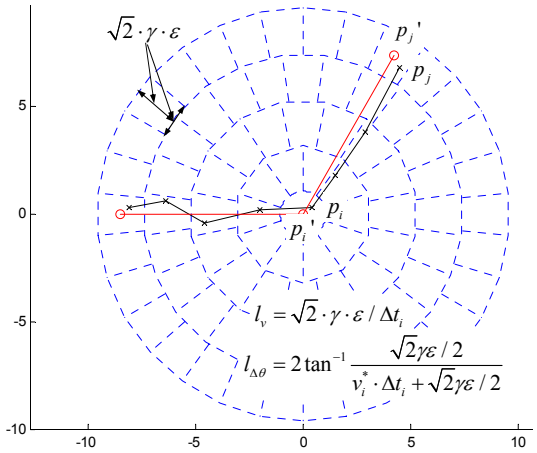


Fig. 1. Example of the approximation process, where polyline $P_i^j$ is approximated by line segment $\overline{p_i' p_j'}$.

estimated as:

$$l_{\Delta\theta}(i,j) = 2\tan^{-1}\frac{\sqrt{2}\gamma\varepsilon/2}{v^*(i,j)\cdot(t_j - t_i) + \sqrt{2}\gamma\varepsilon/2} \quad (3)$$

Thus, the quantized direction change is:

$$\Delta\theta^*(i,j) = [\Delta\theta(i,j) / l_{\Delta\theta}(i,j)] \cdot l_{\Delta\theta}(i,j) \quad (4)$$

### *2.2 Probability estimation*

In the encoding process, we need to encode the quantized value of speed, direction change, while lossless compression is used for time difference.

Adaptive arithmetic coding is applied for encoding the time difference. As the reduction rate may vary in different segments because of the multi-model of the GPS trajectory, a forgetting factor is also used to give a higher weight for recent encoded values. For speed value, its mean and variance are predicted by the previous encoded value in a given time duration, see [12] for more details.

Adaptive arithmetic coding is also used directly for direction change in [12]. However, GPS signals are not always accurate and a quantization step will also cause errors. Therefore, the encoded and true distributions of the direction change are not same. In this paper, Bayes' theorem is applied to improve its probability estimation. Suppose $P(\Delta\theta_0)$ is the distribution of direction change of the clean signal, $P(\Delta\theta_k)$ is the predicted distribution segment $k$, we have:

$$P(\Delta\theta_k) = \sum_b P(\Delta\theta_k \mid \Delta\theta_0 = b) \cdot P(\Delta\theta_0 = b)$$

$$\text{where } P(\Delta\theta_k \mid \Delta\theta_0) \sim N(0, (\rho \cdot \tan^{-1}(r))^2), \quad (5)$$

$$r = (\frac{\gamma^2\varepsilon^2}{6} + \frac{\sigma_{GPS}^2}{2}) / (v_k \cdot \Delta t_k)$$

here $\sigma_{GPS} = 5$ and $\rho = 1.2$ are parameters. $v_k$ and $\Delta t_k$ are the speed and time duration for segment $k$.

After $p_k$ is encoded, posterior probability $P(\Delta\theta_0|\Delta\theta_k)$ is estimated by:

$$P(\Delta\theta_0 \mid \Delta\theta_k) = \frac{P(\Delta\theta_k \mid \Delta\theta_0) \cdot P(\Delta\theta_0)}{P(\Delta\theta_0)} \quad (6)$$

And true distribution $P(\Delta\theta_0)$ is then updated:

$$P(\Delta\theta_0) = \mu_{\Delta\theta} \cdot P(\Delta\theta_0) + P(\Delta\theta_0 \mid \Delta\theta_k) \quad (7)$$

From our experiment, we set 180 levels between $-\pi$ and $\pi$ for $P(\Delta\theta_0)$, $\mu_{\Delta\theta} = 0.995$ is the forgetting factor.

*2.3 Joint optimization process*

In [12], the approximation and encoding process are separated, and a greedy solution is used to get the approximation result first. In this paper, we improve this solution by a joint optimization process.

Suppose $C_t(i, j)$, $C_v(i, j)$ and $C_{\Delta\theta}(i, j)$ are the coding cost for the quantized point $p_j'$ when the previous point is $p_i'$, dynamic programming can be applied by optimizing the following formula recursively:

$$J_j = \min_{\{1 \le i \le j-1\}} (J_i + C_t(i, j) + C_v(i, j) + C_{\Delta\theta}(i, j)) \quad s.t. \delta_{SED}(P_i^j, \overline{p_i' p_j'}) \le \varepsilon \quad (8)$$

As $j - i$ calculations are needed for evaluating the max SED between $p_i$ and $p_j$, the total time complexity of this optimization process will be $O(N^3)$, which is too high for real application. Therefore, a stopping criterion is added to terminate the search when the approximation error is higher than two times of the given tolerance.

Finally, the optimized approximated result with the minimized coding cost can be found by a backtracking process. The pseudo-code can be seen in Algorithm I. Note that the proposed solution can also be used for online purpose directly.

*2.4 Time Complexity*

In the encoding process, the expected time complexity of the proposed algorithm is $O(\tau N^3/M^2)$, where $N$ and $M$ are the number of input and approximated GPS trajectory and $\tau$ is a constant, which is related to the levels in probability estimation. Although the time complexity is slightly higher than GTC algorithm of $O(N^2/M)$, no optimization process is needed in the decoding process and therefore, the same decoding procedure can be applied in $O(\tau M)$ time. Note that the time complexity can be reduced if a hierarchy compression stage is applied with $M \sim N/c$ in each scale, where is $c$ is a constant.

## 3. Experiment and Discussion

In order to evaluate the performance of the proposed Optimized GPS trajectory compression algorithm (OGTC), we use two dataset, Microsoft Geolife dataset with 640 trajectories, 4,526,030 points [13] and MOPSI dataset [1] with 344 trajectories, 744,610 points for testing purpose. These trajectories have a sampling rate between 1s to 5s with different transportation mode such as walking, bus, car, airplane or a multimodal.

The compression performances (KB/hour) are evaluated for different error tolerances: 3m, 10m maximum synchronous Euclidean distance (max SED). The proposed Optimized GPS trajectory compression algorithm (OGTC) is compared with TD-TR + LZMA [3] and GTC [12] algorithm[2]. We can observe in Table 1 that the bit-rate of the proposed algorithm is around 80% compared with GTC algorithm, and it is consistent on both 3m and 10m max SED. Meanwhile, if the original input file is in GPX format, we have a compression ratio around 500:1 on the testing dataset, see Table 2. An example of the proposed compression algorithm can also be seen in Fig. 2.

Note that if a filtering algorithm is performed beforehand, the bit-rate can be reduced around 20% and 15% for 3m, 10m max SED correspondingly. Further information such as proof of the time complexity, details of the experiment result and the matlab code can be seen on http://cs.joensuu.fi/~mchen/GPSTrajComp.htm.

Algorithm I, Approximation and encoding process

**INPUT**
$P = \{p_1, p_2, \ldots, p_n\}$: original trajectory
$\varepsilon$: SED error tolerance
**OUTPUT**
Encoding file

**FOR** $j$ = 2 TO n
  **FOR** $i = j - 1$ TO $i = 1$
    $C(i, j) \leftarrow C_t(i, j) + C_v(i, j) + C_{\Delta\theta}(i, j)$
    **IF** $J_j < J_i + C(i, j)$
      $J_j \leftarrow J_i + C(i, j)$
      $A_j \leftarrow i$ // for backtracking
      Update $P(\Delta t)$ and $P(\Delta\theta^0)$ by (7)
    **ELSEIF** $\delta_{SED}(P_i^j, \overline{p_i' p_j'}) > 2\varepsilon$
      BREAK
    **END**
  **END**
**END**
Backtracking and encoding

[1] http://cs.joensuu.fi/mopsi/

[2] A similar evaluation method is used with commercial software: http://www.droyd.org/gps-trajectory-compression.

Original, 2441 points
193KB for GPX file

3m max SED, 250 points
456 bytes, 0.30 KB/h

10m max SED, 101 points
210 bytes, 0.14KB/h

30m max SED, 46 points
116 bytes, 0.08KB/h

Fig. 2. Compression examples

Table 1. Bit-rate of the proposed algorithm (KB/h)

| | | Geolife Dataset | MOPSI Dataset |
|---|---|---|---|
| 3m max SED | TD-TR[3] | 0.95 | 1.94 |
| | GTC [12] | 0.39 | 0.75 |
| | Proposed | **0.31** | **0.54** |
| 10m max SED | TD-TR | 0.53 | 1.06 |
| | GTC | 0.19 | 0.35 |
| | Proposed | **0.14** | **0.22** |

Table 2. Data Size after Compression (KB)

| | Geolife Dataset | MOPSI Dataset |
|---|---|---|
| GPX file | 436,326 | 77,386 |
| 3m max SED | 1,034 | 155 |
| 10m max SED | 498 | 63 |

## 4. Conclusion

We exploit the problem of lossy compression for GPS trajectories with latitude, longitude and timestamp information, under maximum synchronous Euclidean distance (max SED). Dynamic programming is used to seek an optimized approximation result with the minimized coding cost. The prediction and estimation of direction change is also improved by Bayes' theorem.

Experimental results show that the proposed method achieves 0.31 and 0.54 KB/h in Microsoft Geolife dataset and MOPSI dataset for 3m SED, around 80% bit-rate comparing with the previous GTC algorithm, while the decoding time will not increase. For GPX file, the proposed algorithm achieves around 500:1 compression ratio with 3 meters accuracy.

## Acknowledgement

This work is supported by Tekniikan edistämissäätiö (TES), Nokia Scholarship, MOPSI project EU (EAKR), National Natural Science Foundation of China (Grant No 61072146) and Shanghai Committee of Science and Technology, China (Grant No 10PJ1404400).

## References

[1] Z. Yu, X. Zhou, *Computing with Spatial Trajectories*, Springer, 2011.

[2] D. H. Douglas, T. K. Peucker, "Algorithm for the reduction of the number of points required to represent a line or its caricature", *The Canadian Cartographer*, 10 (2), 112-122, 1973.

[3] N. Meratnia, R. A. de By, "Spatiotemporal Compression Techniques for Moving Point Objects", *Proceedings of the Extending Database Technology*, 765-782, 2004.

[4] M. Potamias, K. Patroumpas, T. Sellis, "Sampling Trajectory Streams with Spatiotemporal Criteria", *Scientific and Statistical Database Management*, 275-284, 2006.

[5] M. Chen, M. Xu, P. Franti, "A Fast *O*(*N*) Multi-resolution Polygonal Approximation Algorithm for GPS Trajectory Simplification", *IEEE Trans. on Image Processing*, 2012.

[6] F. Schmid, K. F. Richter and P. Laube, "Semantic Trajectory Compression", *Lecture Notes in Computer Science*, vol. 5644, 411-416, 2009.

[7] G. Kellaris, N. Pelekis and Y. Theodoridis, "Trajectory Compression under Network Constraints", *Lecture Notes in Computer Science*, vol. 5644, 392-398, 2009.

[8] J. Muckell, J. H. Hwang, C. T. Lawson, S. S. Ravi, "Algorithms for compressing GPS trajectory data: an empirical evaluation", *SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 402-405, 2010.

[9] A. Kolesnikov, "Fast algorithm for error-bounded compression of digital curves", *IEEE Int. Conf. on Image Processing*, Hong Kong, China, 1453-1456, 2010.

[10] M. Chen, M. Xu, P. Franti, "Fast dynamic quantization algorithm for vector map compression", *IEEE Int. Conf. on Image Processing*, Hong Kong, China, 4289-4292, 2010.

[11] M. Koegel, M. Mauve, "On the Spatio-Temporal Information Content and Arithmetic Coding of Discrete Trajectories", *International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, Copenhagen, Denmark, December, 2011.

[12] M. Chen, M. Xu, P. Fränti, "Compression of GPS Trajectories", *Data Compression Conference*, 62-71, Snowbird, USA, 2012.

[13] Y. Chen, K. Jiang, Y. Zheng, C. Li, N. Yu, "Trajectory Simplification Method for Location-Based Social Networking Services", *ACM GIS workshop on Location-based social networking services*, 33-40, 2009.