MINING SPATIAL DATA FROM GPS TRACES FOR AUTOMATIC ROAD NETWORK EXTRACTION

F. Lima[a, *], M. Ferreira[a, b]

[a] Artificial Intelligence and Computer Science Laboratory, University of Porto, Portugal - (flins,michel)@dcc.fc.up.pt
[b] Department of Computer Science, University of Porto, Portugal

KEY WORDS: Data Mining; Database; Extraction; GIS; GPS; Mapping.

ABSTRACT:

The car manufacturing industry has been conducting a considerable effort to allow future vehicles to communicate, either between them or with a road infrastructure, in order to improve driving safety. As the position of each vehicle is an essential attribute of the proposed application protocols (to avoid collisions at blind intersections, for instance), and is also fundamental to support complex network protocols based on mobile wireless nodes with very limited transmission range, such communicating vehicles will be further equipped with GPS receivers. This massive distribution of GPS sensors, in conjunction with a free of charge communication infrastructure that allows accessing the information collected by such devices, will create a powerful new medium of remote sensing of geographical information. In this paper we address the automatic road network extraction based on this vehicular sensing infrastructure where the sensor in play is just the GPS receiver. We have resorted to the widely available GPS/GPRS tracking technology, heavily used by trucking companies, in order to obtain more than 30 million GPS points to construct the road map of an interesting city of Portugal, called Arganil, in an accurate, inexpensive and permanently up-to-date manner. Our algorithm is implemented using spatial SQL queries to aggregate data from multiple traces to produce a weighted-mean geometry of road axles, diluting GPS errors. In order to evaluate our extracted road network, we have compared its geometric and topological layers with a vectorial road map extracted from high resolution satellite images. Results show a highly accurate correspondence between them in all areas where a sufficient number of GPS traces have been collected.

1. INTRODUCTION

Spatial data acquisition is one of the most expensive and time-consuming tasks in the deployment and updating of Geographic Information Systems (GIS). A few decades ago, topographic surveys and aerial images interpretation were the main sources of spatial information for cartographic purposes. Because of the complexity of such processes, cartographic updating was not done very often, but only when specific needs or major landscape changes mandated the acquisition and processing of new spatial data. More recently, the constant advances in Remote Sensing have allowed a fast-paced production of large volumes of spatial information. Several remote sensors orbiting Earth are now able to easily acquire images suitable for monitoring the intra-urban landscape changes, as those aboard satellites such as SPOT-5, CBERS-2b, IKONOS II, QuickBird 2, OrbView and Eros [1].

An alternative to such remote sensing technologies is emerging in the form of standard on-site mobile sensors, supported by the ubiquitous deployment of positioning technologies, such as the Global Positioning System (GPS). GPS devices are capable of assigning a geographic position to every type of data collected by on-site sensors. The wide popularity of GPS-based in-car navigation systems, together with the typical myriad of sensors installed in modern vehicles, is creating a powerful new infrastructure for the acquisition of spatial data through a distributed vehicular network.

Motivated by the improvement of driving safety, the automotive industry has been installing a variety of sensors in modern production vehicles, capable of providing several data regarding the vehicular behaviour, as well as data related to the environment where the vehicle moves. In addition, the car manufacturing industry is conducting a major effort towards the development of a distributed, peer-to-peer and infrastrutureless communication network composed of vehicles and based on Digital Short-Range Communications (DSRC) [2]. Such network is being defined to allow both vehicle-to-vehicle and vehicle-to-road-infrastructure communication. The combination of a massive distribution of positioning sensors and a charge-free communication mean that allows accessing the information collected by vehicular sensors is thus able to create a novel and powerful tool for the acquisition of spatial data. Based on the remote sensing of a large network of vehicles, it is possible to build detailed maps of highly dynamic geographic information, related to the road network, such as traffic conditions, temperature and wetness of road pavement, air pollution levels, or even the presence of potholes on the road.

Much less dynamic, but also interesting, is the information that allows mapping and updating the actual road network based on vehicular infrastructure of remote sensing. The goal is not only the acquisition of road axle geometries, but also their characterization in terms of topological connectivity, traffic rules and speed patterns, in an accurate and permanently up-to-date manner.

Several projects have been developed with the goal of making a better use of the data collected through GPS receivers [3-7]. One of the most important of these projects is the OpenStreetMap that hosts a collaborative network of GPS traces for the assisted construction of road maps. Despite the increasing research around this area, very few references relax the need of a base map in a non-assistive approach [3,4]. Previous work has focused more on refinement issues and updating of existing cartography.

* Corresponding author.

In this paper we address the automatic road network extraction, where the sensor in play is just the GPS receiver, based on millions of points collected by a fleet of vehicles. Given the non-existence today of the described vehicular networks, we resorted to the widely available GPS/GPRS tracking technology, heavily used by trucking companies. The constant bandwidth increase in cellular networks is allowing such tracking technology to transmit position reports "in-raw", i.e., as received by the on-board GPS unit, with a point every second. The automatic road network extraction algorithm is implemented using spatial SQL queries to aggregate data from multiple traces to produce a weighted-mean geometry of road axles, diluting GPS errors. It does not require a base map and any editing from users. The paper is organized as follows: the next section describes the process of collecting and filtering GPS data. Section 3 presents the road network extraction algorithm. Section 4 reports the experimental results. Section 5 introduces a technique for inferring road classification. We end by outlining some conclusions.

2. DATA COLLECTION AND FILTERING

The automatic construction of road map from GPS traces requires the availability of a large data set collected over the area of interest. In our implementation, we used more than 30 millions of GPS points, collected in real-time by a vehicle tracking company, using a temporal detail of one point per second. Such level of detail is particularly important for the representation of the road network, since it allows keeping the geometry continuity of the vehicular trajectories.

The position of each point is defined by its geographic coordinates: longitude, latitude and altitude. To meet our purpose of accurate road map construction, we extended the protocol to also include information about the number of satellites and the horizontal dilution of precision (HDOP). We also stored additional relevant information, such as speed of the vehicle, its azimuth and the time of the position reading. We collected a total of 371600 km of vehicular traces, spatially distributed in Portugal. Our implementation was tested in the city of Arganil (Fig.1).

Figure 1. GPS Traces on Arganil

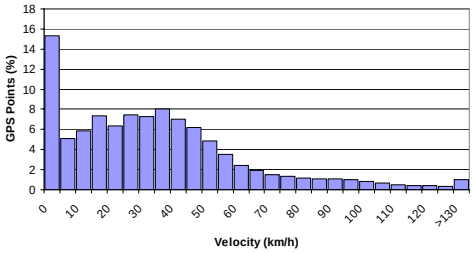

Figure 2. Speed distribution

Because of GPS errors, it is necessary to rely on several processes that allow the elimination of inconsistent data, aiming at obtaining higher quality input data to our algorithm. We have thus established three filters, where the first one is based on speed information. Points collected at speeds lower than 6 km/h were not considered to be sufficiently accurate for the automatic construction of road maps. As a result of this filter, we have eliminated 15,31% of the collected points. Figure 2 shows the speed distribution on our data set.

Our second filter is based on the HDOP value, which is a measure quantifying the degradation level of the horizontal positioning accuracy of the GPS (2D-based positioning). This value is mainly determined by the relative geometry of the visible satellites when the positioning reading was taken. A low HDOP value means a more accurate horizontal positioning. The distribution of HDOP values in our data set is shown in Fig. 3. The filter has been configured to eliminated points with an HDOP value higher than 2.

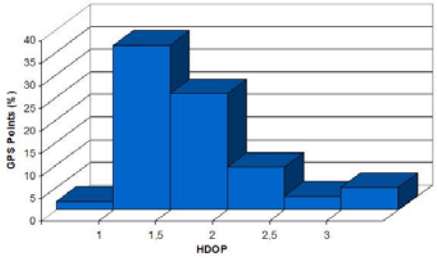

Figure 3. HDOP distribution

Our third filter in the pre-processing phase of our data is based on the number of satellite used in the positioning. When more than 4 satellites are used for the positioning, the redundant satellites can serve for the detection of erroneous readings, thus increasing the accuracy of the positioning. The distribution of the number of satellites used to obtain a positioning in our data set is shown in Fig.4. The filter is set to eliminate points collected using a number of satellites lower than 5.

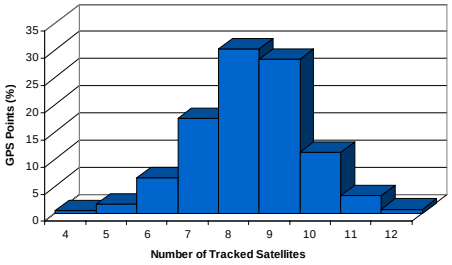

Figure 4. Tracked satellites distribution

In addition to the filters mentioned above, the pre-processing phase has an extra module that is responsible for the

identification of large intervals of time between consecutive points of the same vehicular trace, which are either caused by obstacles to the reception of the signal broadcasted by the satellites, or by the elimination of points from previous filters. Such large intervals can erroneously affect the geometry of the road network and we thus divide vehicular traces where two consecutive points are separated by more than 7 seconds into two distinct traces.

The last step in our pre-processing phase consists in the simplification of the GPS traces in order to cope with performance issues of our algorithm and minimize the amount of memory required to store the vehicular traces.

Our traces have been simplified using the Douglas-Peucker algorithm [8], resulting in the elimination of 67% of the collected points (Table 1). The maximum distance allowed for the elimination of point through this algorithm was of 1 meters.

	Original traces	Simplified traces	Elimination rate
Total number of points	35041489	11628278	67%
kms	511067,34	510740,69	0,06%

Table 1. Line Simplification by Douglas-Peucker algorithm.

3. ROAD NETWORK EXTRACTION

The main goal of our algorithm is the construction of a graph representing the road network, where the roads are represented by edges and the intersections are represented by nodes. The algorithm is implemented through spatial SQL queries to aggregate data from multiple traces, in order to produce an weighted geometry of road axles, diluting errors from the GPS receivers. The input data is stored in two tables of a spatial database: traces and points.

The algorithm is divided in five steps: rasterization, centroid generation, geometric connectivity of the centroids, topologic connectivity (node-edge topology) and turn-table construction. We next describe each of these steps.

3.1 Rasterization

Until May, 2000, the real-time positioning of a point, through a navigation GPS receiver, provided a planimetric accuracy better than 100 meters. Since then, with the ending of the Selective Availability (technique used to degrade the accuracy of the positioning), such value became, on average, better than 15 meters. Even with this significant improvement, the attained accuracy is not considered to be sufficient for a valid geometric representation of the road network. Aimed at an accurate and automatic construction of road maps, we propose the spatial aggregation of a large set of GPS traces through a rasterization process.

The term rasterization is used in the context of the transformation of a vectorial representation into a matrix-based representation. In the work described in this paper, the rasterization process enables the transformation of the vectorial layer of GPS traces into a raster layer of 5-meter-resolution cells. For each cell, we assign a value that translates the number of GPS traces that intersect it, as returned by the following SQL query:

```
SELECT m.id, COUNT(t.id) FROM traces t, matrix m WHERE
ST_Intersects(t.trace,m.pixel) GROUP BY m.id
```

Figure 5. Rasterization process: cells represented by an unique symbol

Figure 5 presents the set of cells intersected by one or more GPS traces. Such set represents, in a fuzzy manner, the roads travelled by the vehicular sensors, resulting in blurred areas in places where the road network is especially complex. However, if we use the attribute that holds the number of traces intersecting each of the cells, to vary its color intensity (depicted using different level of gray), it becomes possible to easily identify the road axles (Fig.6)

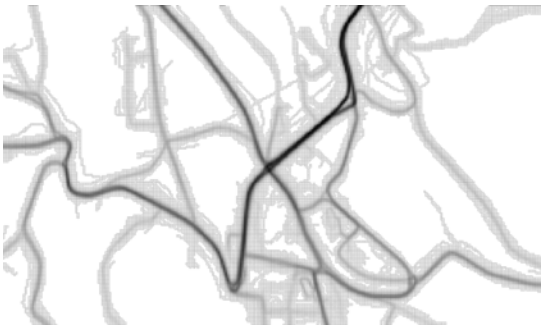

Figure 6. Rasterization process: cells represented by a gray scale based on their values

The probability of existing a road in a given cell is proportional to the value of the attribute of the cell. Similarly, cells holding a low counting value of intersecting traces represent disperse vehicular trajectories or low-travelled roads. The rasterization step thus performs an highly refined filtering of our data set, using a process based on spatial aggregation together with sampling correction. This approach becomes particularly important to the representation of small roundabouts, nearby roads and other complex parts of the road network.

Furthermore, the rasterization process allows a better representation of wide roads (e.g. roads with two ways separated by a central structure), becoming possible the identification of a road axle in each of the directions (Fig.7).

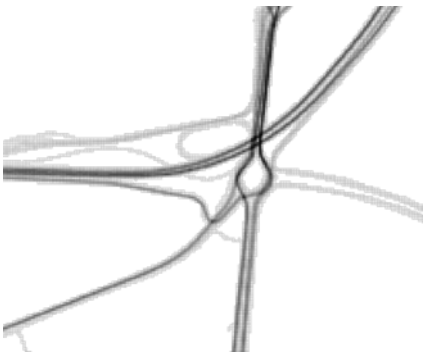

Figure 7. Identification of road axles in wide roads

3.2 Centroid Generation

Centroids can be defined as the points belonging to the axle of the road. The first step for its generation is the descendent ordering of the cells generated in the rasterization process, according to the attribute value of each. As explained previously, the higher the value of a cell, the higher the probability of existing a road on it. In our implementation, only cells presenting a value higher than 20, i.e. cells with at least 20 vehicular traces intersecting it, are classified as cells generating centroids. After identifying such cells, the position of candidate (x_{cent},y_{cent}) is given by:

$$x_{cent} = \Sigma (x_i * v_i) / \Sigma v_i$$
$$y_{cent} = \Sigma (y_i * v_i) / \Sigma v_i \quad (1)$$

where x_i, y_i = geographic coordinates of the center of mass of the cell with higher value and its adjacent cells;
v_i = attribute value of such cells.

The candidate centroids are only added to the database if there is no previous centroid within a distance d_i of them. The distance allowed for the addition of a candidate centroid is related to the attribute value of the cell to which the candidate centroid belongs. The higher the value, the lesser distance between neighbour centroids. Such constraint is designed to minimize the zigzag effect between centroids representing the same road axle. Table 2 presents the correlation between the number of cells generated in the rasterization process and the total number of centroids generated.

Total of cells (value > = 1)	Generating cells (value >20)	Generated centroids	Rate
1.118.821	266860	32310	12,11%

Table 2. Centroids Generation.

The output of the second step of our algorithm is shown in figure 8.

Figure 8. Centroids Generation

3.3 Geometric Connectivity

The third step in our algorithm deals with the establishment of geometric connectivity between centroids, according to the following spatial query:

```
SELECT c1.id,c3.id FROM centroides c1,centroides c2, centroides c3 WHERE
c2.id=id AND c1.id<>c3.id AND c1.id<>c2.id AND c3.id<>c2.id AND
ST_DWithin(c1.centroide,c2.centroide,dist) AND
ST_DWithin(c3.centroide,c2.centroide,dist) AND EXISTS (SELECT t.id from
trajetos t WHERE ST_DWithin(trajeto,c1.centroide,1.5) AND
ST_DWithin(trajeto,c2.centroide,1.5) AND ST_DWithin(trajeto,c3.centroide,1.5)
AND ST_line_locate_point(trajeto,c2.centroide) BETWEEN
ST_line_locate_point(trajeto,c1.centroide) AND
ST_line_locate_point(trajeto,c3.centroide)) ORDER BY
ST_Distance(c1.centroide,c2.centroide), ST_Distance(c3.centroide,c2.centroide)
```

According to this query, the connection between adjacent centroids is allowed when one or more GPS traces pass near them, with a maximum distance of 1.5 meters. This step leads to the geometric representation of a road (Fig.9).

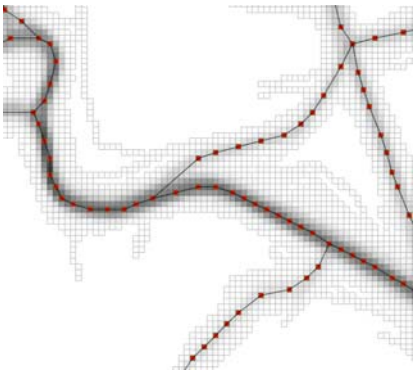

Figure 9. Geometric connectivity between adjacent centroids.

3.4 Topological Connectivity

The fourth step in our algorithm addresses the topological connectivity of the road network (edge-node topology). For this step it is essential that the road network representation is consistent with the network model, requiring that intersections are represented by nodes and road axles (connecting neighbor nodes) are represent by edges.
The identification of nodes is done through a spatial query that derives a list of centroids located at the end points of at least three distinct segments. Such centroids are then classified as nodes of the road network being constructed. We then connect the existing segments between two neighbour nodes, deriving the set of edges of our network model. Figure 11 illustrates the results of this step.

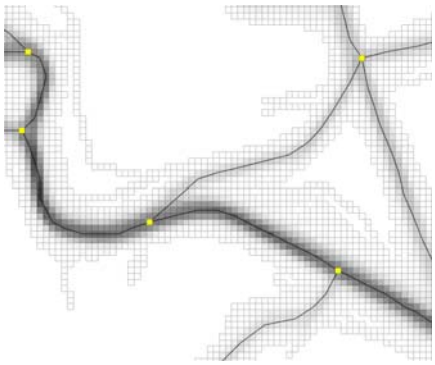

Figure 10. Topological connectivity: road network as a set of nodes and links

3.5 Turn Table Construction

Topological connectivity is the key element in a transportation network, since it determines the mobility patterns in such a network. However, the typical edge-node model can include nodes which do not necessarily represent a real intersection (typically in scenarios with three dimensional connectivity through elevated roads); or nodes that geometrically connect uni-directional edges, where traffic rules disallow specific

maneuvers from one edge to another over such geometric connectivity. The fifth step in our algorithm of automatic road map construction addresses the derivation of the turn table, as mandated by traffic rules.

To construct the turn table, we implemented an adapted module of *map-matching* that uses the GPS traces to extract traffic rules of inter-edge connectivity. The turn table consists of two attributes: the identifier of the source edge; and the identifier of the destination edge. Figure 11 presents the final structure of our spatial database supporting the algorithm described in this paper.

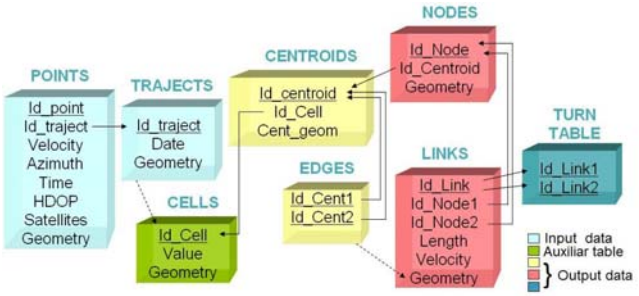

Figure 11. Spatial Database structure

4. EXPERIMENTAL RESULTS

Using the algorithm for automatic road network extraction, we produced a vectorial road map of the municipality of Arganil. The results show an accurate overlap of the extracted road network with that from Google Maps, in every place where a sufficient number of GPS traces has been collected (Fig.12).

Figure 12. Overlapping between the extracted road network and Google Maps

The evaluation phase of the results obtained was performed by comparing the geometric and topological layers of the extracted road network with those from vectorial maps provided by the map-making company InfoPortugal, S.A., which are constructed by manually processing orthorectified aerial images, with a resolution of 25cm² per pixel.

Coverage evaluation is done through three main metrics: total number of kilometers of roads generated by the algorithm in the zone of relevance (Table 3); total number of kilometers of roads to which a match is found in InfoPortugal's map; total number of kilometers of roads that are not present on InfoPortugal's map (cartographic updating).

	InfoPortugal's map	Extracted road network	Percentage
Total of kms	522,778	421,82	80,69%

Table 3. Extracted road network statistics

The association between the extracted road axles and the existing road axles in InfoPortugal's map (Table 4) is determined through a map matching algorithm.

Figure 13 shows the correspondence between the extracted road network (in black color) and InfoPortugal's map (in red color).

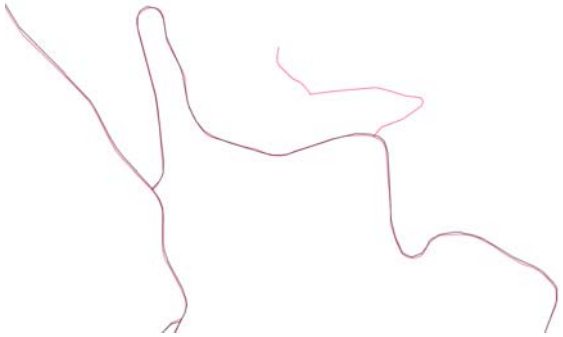

Figure 13. Correspondence between the extracted road network, in the municipality of Arganil, and InfoPortugal's map.

	Total of km	Correspondence (km)	Correspondence Rate
Extracted road network	421,82	349,87	82,94%

Table 4. Correspondence between the extracted roads and InfoPortugal's map

One of the most important results from the process of automatic road network extraction based on vehicular GPS traces consists in the identification of roads that are still non-existent in current maps, as shown in Fig. 14. This aspect shows the ability of the algorithm to provide an inexpensive and highly accurate way of constantly performing cartographic updating.

Table 4 presents a correspondence of 82,94% between the kilometers of roads that were extracted and those from the base map. Hence, the remaining 17,06% represent the percentage of cartographic updating, i.e., the extraction of non-existing roads in the base map.

Figure 14. Cartographic Updating

The evaluation of the accuracy of the geometry of the extracted road network is done in a continuous manner: the average distance between the extracted roads and the corresponding roads in the base map is obtained through the computation of

the area between the two, divided be the average length of the two geometric representations (Fig.15). This method presents a more realistic evaluation as compared to discrete based measurements.

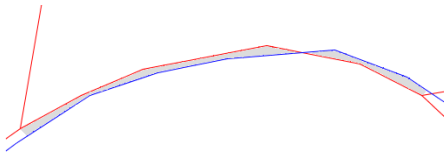

Figure 15. Evaluation of geometric accuracy: area between two representations of the same road

The resulting average distance was of 1,43 meters between the two representations of the same roads.

The evaluation of the topology of the extracted road network is done through two main metrics: the number of nodes that have been extracted that match nodes in the base map; and the traffic direction of each extracted edge, compared to the direction of the associated edge in the base map (Tables 5 and 6).

Extracted nodes	Matching nodes	Percentage
596	458	77%

Table 5. Correspondence between extracted and base map nodes

The nodes which have no matching node on the base map do not necessarily represent false intersections. Such nodes can result from junctions of wide roads into more narrow segments, where the algorithm of automatic extraction usually created a non-existent node in the base map. Such nodes can also result from cartographic updating and the identification of intersections with non-existent road in the base map.

Number of tested links	Correctly classified links	Rate of correct classification
1206	1158	96%

Table 6. Edge direction

The last evaluation concerns the traffic rules between the generated roads. The extracted enforcements between connectivity of specific edges, stored in the direction table, were checked against the Do Not Enter signs of the base map (table 7).

Do Not Enter signs	Inconsistent traffic rules in the direction table	Average error
50	1	0,02 %

Table 7. Traffic rules evaluation

5. CONCLUSIONS

In this paper we have presented an algorithmic methodology for the automatic extraction of road networks based on vehicular GPS traces. The current availability of GPS receivers for navigation allows a large-scale acquisition of spatial data related to the road network. The spatial aggregation of information produced by multiple GPS traces allows the construction of an weighted mean geometry of roads, followed by the determination of the respective topological connectivity and the identification of traffic rules associated to the extracted road segments.

The methodology we presented has some relevant advantages as compared to the traditional approach based on road network extraction by the manual processing of high-resolution aerial images. Such advantages are translated essentially in the simplification of the process of cartography updating, that is able to be done in an accurate, inexpensive and frequent manner. Furthermore, even through the use of commercial GPS receivers, which present an average accuracy of 15 meters, it is possible to extract the geometry of roads with an accuracy better than 1.5 meters, making use of the aggregation of a large amount of readings. The rasterization process, explained in section 3, works as a natural filter, eliminating outliers and enabling the accurate determination for the position of centroids, resulting in an higher accuracy for the geometry of the extracted road network. Another advantage of this approach is the ability to detect roads which are not visible from aerial images, because of the presence of clouds, tree of shadows. In addition, the use of a large number of GPS traces makes it possible to extract the geometric representation of detailed features of the road, such as small roundabouts ou nearby parallel roads, without having to resort to much more expensive devices, such as differential GPSs.

Our approach deals not only with the geometry of roads, but also with the associated characterization in terms of traffic rules and speed patterns, in an accurate and always up-to-date manner. Such characterization cannot be obtained by the processing of satellite or air-Bourne images.

The existence of vehicles equipped with a variety of sensors, together with a large-scale distribution of GPS receivers and the development of an inter-vehicular ad hoc network, will enable the arrival of a novel and powerful remote sensing infrastructure, suited to provide highly detailed geographic information about and around the space of roads. This advent is creating an interesting and rapidly expanding area of research, built around the theme of vehicular sensing.

Acknowledgments. This work has been partially supported by the Portuguese Foundation for Science and Tecnology (FCT) under project JEDI (PTDC/EIA/66924/2006) and by funds granted to LIACC through the Programa de Financiamento Plurianual and POSC.

REFERENCES

1. Farina, F. C., Ahlert, S., Duranti, R. R., Silva, T. P., Fagundes, C. L., 2007. Utilização de imagem de alta resolução espacial para o mapeamento do município de Monte Belo do Sul, RS. In: *XIII Simpósio Brasileiro de Sensoriamento Remoto*, Florianópolis, Brasil, INPE, pp. 515-521.
2. Bera, R., Bera, J., Sil, S., Dogra, S., Sinha, N.B., Mondal, D., 2006. Dedicated short range communications (DSRC) for intelligent transport system. In: *Wireless and Optical Communications Networks IFIP International Conference*, 0(0), pp.0 – 5.
3. Brüntrup, R., Edelkamp, S., Jabbar, S., Scholz, B., 2005. Incremental Map Generation with GPS Traces. In: *Proc. of the 8th Int'l. IEEE Conf. on Intelligent Transportation Systems*, Vienna, Austria, pp. 574 – 579.
4. Edelkamp, S., Sulewski, D., Pereira, F. C., Costa, H., 2008. Collaborative Map Generation - Survey and Architecture Proposal. *Urbanism on track.*

5. Davies, J. J., Beresford, A. R., Hopper, A., 2006. Scalable, Distributed, Real-Time Map Generation. *IEEE Pervasive Computing*, 5 (4), pp. 47-54.
6. Rogers, S., Langley, P., Wilson, C., 1999. Mining GPS data to augment road models. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, pp. 104-113.
7. Schroedl, S., Wagstaff, K., Rogers, S., Langley, P., Wilson, C., 2004. Mining GPS Traces for Map Refinement. *Data Mining and Knowledge Discovery*, 9, pp. 59-87.
8. Douglas, D. H., Peucker, T .K., 1973. Algorithms for the Reduction of the Number of Points Required to Represent a Line or Its Caricature. *The Canadian Cartographer*, 10(2), pp.112–122.