INCREMENTAL DATA ACQUISITION FROM GPS-TRACES

L. Zhang and M. Sester

Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover Germany- (Lijuan.zhang, monika.sester)@ikg.uni-hannover.de

Commission IV, WG IV/2

KEY WORDS: Integration, Incremental refinement, GPS data, Road map

ABSTRACT:

GPS traces can track actual time and coordinates of regular vehicles going their own business, and it is easy to scale to the entire area with an accuracy of 6 to 10 meters in normal condition. As a result, extracting road map from GPS traces could be an alternative way to traditional way of road map generation.

The basic idea of this paper is to describe a process which incrementally improves existing road data with incoming new information in terms of GPS traces. In this way we consider the GPS traces as measurements which represent a "digitization" of the true road. Although the accuracy of the traces is not too high, due to the high number of measurements an improvement of the quality of the road information can be achieved.

Thus, this paper presents a method for integrating GPS traces and an existing road map towards a more accurate, up-to-data and detailed road map. First we profile the existing road by a sequence of perpendicular profiles and get the road's candidate sampling traces which intersect with the profile. Then we match the potential traces with the road and finally estimate the new road centerline from its corresponding traces. In addition to the geometry of roads we also mine attribute information from GPS traces, such as number of lanes. Furthermore, we explore the benefit of an incremental acquisition by a temporal analysis of the data.

1. INTRODUCTION

Nowadays, GPS data are becoming more and more available: most cars are equipped with low cost GPS receivers, which potentially accumulate a lot of data just while driving. The data have an accuracy of 6 to 10 meters in normal condition (Haklay and Weber, 2008). All the above make it possible to generate road map from GPS data. Compared with the traditional way of road map generation it has the following benefits: low cost and particularly it can keep up with changes.

A lot of projects attempt to make use of these valuable data. One of the most extensive and effective perhaps is the OpenStreetMap project (Haklay and Weber, 2008). The project aims to create a set of map data that is free to use and editable, since accurate digital geographical information is considered to be expensive and out of the reach of individuals, small businesses, and community organizations. It has an increasing number of volunteers to contribute to the project and has gathered a large volume of GPS data. The users of the project can edit the map with GPS traces, out of copyright maps and satellite images manually.

As open street map can provide the GPS data, we would like to eliminate the manual step and generate road maps from GPS data automatically, and get a more accurate, detailed and up-to-date road map. We also try to mine attribute information from GPS traces, such as number of lanes. Furthermore, we explore the possibility to also derive time dependent phenomena such as temporal blockage of a road or varying usage of roads over the day, week or year. In this way, incrementally the existing information can be improved with each incoming new data set. However, as our example data sets do not contain absolute time stamps, the temporal analysis in this paper is restricted to a theoretical description.

2. RELATED WORK

In recent years new data sources are being available like massive amounts of data collected by volunteers (Goodchild, 2007) like GPS-traces of hikers or car-drivers, which in principle are manifestations of digitizations of roads or footpaths. The integration of GPS tracks mainly has to deal with the high degree of noise resulting from the low quality of the GPS measurements. This makes it on the one hand difficult to discern nearby roads and on the other hand also to reconstruct the underlying structure in the road geometry, e.g. the number of lanes. In order to derive an integrated geometry from the collection of given tracks, aspects of reliability and trust (Sayda, 2005) as well as geometric accuracy have to be taken into account.

In case of the road tracks, the goal is to reconstruct the centerline as well as the number of lanes from the noisy road data. Most of the approaches use histograms in profiles orthogonal to the hypothesized road. The mean of the intersection points of the profile with the traces delivers points of the centerline of the road. In order to separate different lanes, Schroedl et al. (2004) propose to find clusters in dedicated distances, corresponding to the typical width of lanes. Cao & Krumm (2009) use an approach based on a force model which optimizes the displacement of individual tracks towards a

modeled center line. Chen & Krumm [2010] consider the distribution of tracks on the different lanes as a mixture of Gaussians and therefore use a Gaussian mixture model (GMM) to model the distribution of GPS traces across multiple lanes; also here prior information about lane width and corresponding uncertainty is introduced.

Davies et al. (2003) use a raster-based approach – similar to occupancy grids used in robotics – in order to determine the geometry of roads. In their approach, they also have a temporal component by including a kind of fading of roads which are not regularly frequented. In this way, abandoned roads can be identified. Thus they are able to also describe the temporal change of objects.

The approach presented in this paper also uses a clustering approach. In contrast to existing work, the distinction of different roads also takes the velocity of the tracks into account. In this way, especially highway exits can be discerned from the highways themselves. Furthermore, we propose to exploit the sequence in the trajectories to infer temporal attributes about the road.

3. DATA SETS AND PREPROCESSING

3.1 Data Sets

GPS data can be downloaded from the OpenStreetMap website. The GPS data are recorded and contributed by OpenStreetMap users doing their own business. The source of the data can be from cars, pedestrians and bicycle riders. The typical accuracy of the data is 6 to 10 meters in normal conditions. The GPS data are not distributed equally among the roads. Some roads have more corresponding traces than others. In our research area, we learn that a typical highway has 30 to 80 corresponding traces, whereas a busy city road has less than 20 traces and a road in a local neighbourhood has none or only a few. Even when roads are of the same class and close to each other, the number of their corresponding traces may vary noticeably.

Besides the raw GPS-data, OSM mainly contains edited road data sets, which correspond to “the true road object”. Since the map can be changed by anyone the map has not been checked or verified. The road map has the attribute “ONEWAY” indicating whether the road is a one-way road or not. We use this road data set as reference data to start the search for corresponding GPS-traces. Furthermore, we used TeleAtlas-data for an independent quality analysis.

Unfortunately, the data sets did only contain relative time stamps, but no absolute ones. Therefore, no experiments concerning the temporal patterns could be conducted.

3.2 Preprocessing

The data sets are preprocessed before we do the integration. The data sets consist of individual GPS points, which have latitude, longitude and sometimes a time stamp. GPS points are linked according to time sequence. In some cases there are unreasonable links between different trips. Therefore, we have to split GPS trace into individual trips. We split the trace whenever the distance between two points is larger than 300 meters or the change of direction is larger than 45 degree.

We also derive the speed of the traces from the GPS data. Most of the GPS points are recorded with the time interval of 1 second, but due to the diversity of loggers, the interval between GPS points are not the same for all points and it can be a few seconds for some points. In highway area we set 250 kilometers per hour as the limit of the vehicle‘s speed: if the speed for one line segment is larger than 250 kilometers per hour we just think that it is the distance the vehicle moved in 2 or more seconds and calculate the speed till it is less than 250 kilometers per hour. In the urban area we use 100 kilometers per hour as a threshold to calculate the speed of the traces. Figure 1 shows a section of the data set in the highway area before and after preprocessing. Figure 2 shows the speed of the traces in two areas: darker color indicates lower speed. It clearly shows the lower speeds in the exit lanes of the highway situation, and also on the left-turn lane in the inner-city situation.

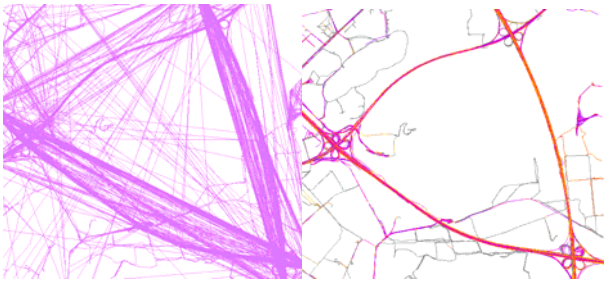
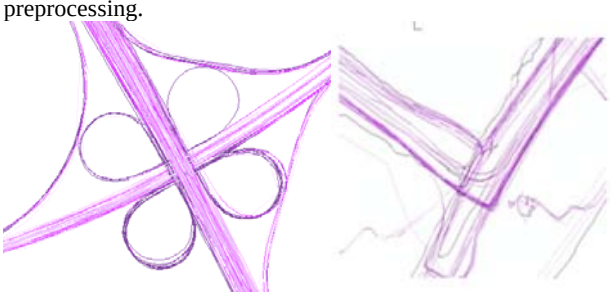


Figure 1: GPS traces in highway area, before and after preprocessing.



(a) Highway area (b) Urban area

Figure 2: Traces are shown in different shades of violet representing different speed ranges in highway area and urban area. Darker color indicates lower velocity.

4. EXTRACTING ROAD CENTERLINE

The challenge in interpreting and integrating the GPS-traces is firstly to determine the centerline from multiple representatives of GPS traces. Furthermore, if several roads are nearby, they have to be separated appropriately. We consider the individual GPS-traces as measurements which are associated with a certain error. The “true” geometry is then derived by averaging all traces corresponding to one road. In order to start the process, we use the reference road map from OSM as initial prior information. In order to determine the road centerline, we sample it at certain distances, by putting profiles perpendicular to the initial road. The intersections of the profile with the GPS-traces deliver sampling points for the road centerline. The whole process for the extraction of the road centerline is visualized in Figure 3.

4.1 Matching Method

The prior road map uses sequences of line segments that connect coordinate points which represent the centerline geometry. If a road’s “ONEWAY” attribute is yes, the road has a direction that accords with the sequence of its line segment. Otherwise, the sequence of line segment does not indicate the

road's direction. We then say the road has no direction and it means that the vehicles can drive in both directions on it.

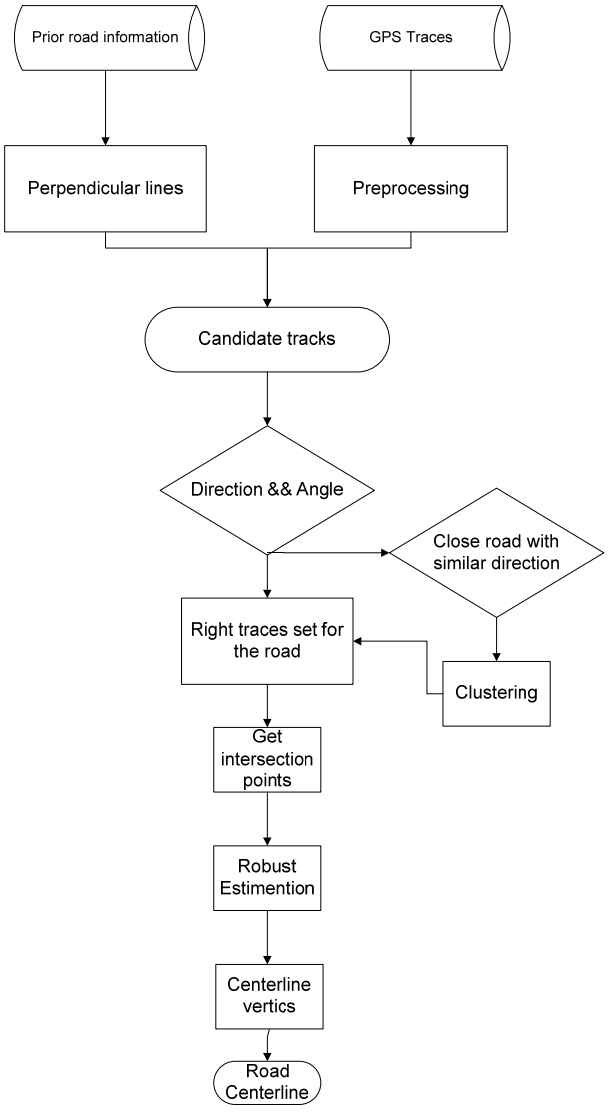


Figure 3: Work flow for extracting of road centerline.

There are three conditions we used to find corresponding traces to a priori road: distance to the road, direction, the angle between the trace and road. First, as shown in Figure 4, we determine profiles along the road and with a width of 30 meters. We try to use wide enough profiles to make sure that all possible traces for the road are included. Since the error of GPS traces can reach 10 meters, we try 10 meters, 20 meters, and 30 meters. We find that 30 meters buffer is suitable to select possible traces. The profiles are perpendicular to the line segment's direction that they belong to. The traces that intersect with the profile are candidate traces for the road. Second, traces are removed from candidate traces set if the angle between them and the road is larger than 20 degrees. Here we also make experiments to make sure that the angle threshold is neither too small to neglect right traces nor too large to select wrong traces. At last, if the road has a direction, only those traces having the same direction as prior road remain in the candidate traces set. Using this matching method, the traces can be assigned to right road if there is no neighboring road that is close enough and has similar direction. However, if the situation happens, traces cannot be separated from its neighborhood road's traces. Figure 4 shows such a case where the circular road is close to the straight road. In order to separate also such cases, in addition to the above measures we use a clustering method to separate the traces. The clustering also takes the difference of velocity of the tracks into consideration.

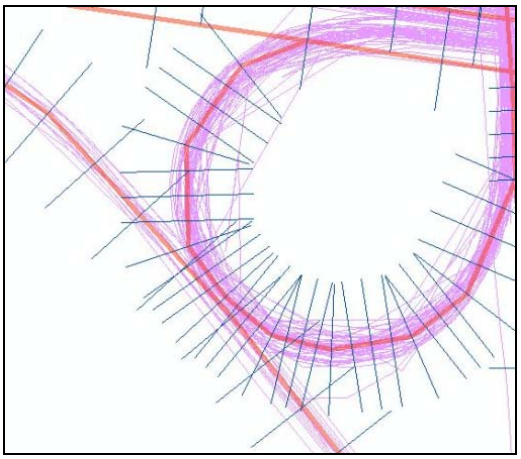


Figure 4: Getting candidate traces for the road using p perpendicular (in green) to prior road's centerline (in red).

4.2 Separate traces when two roads are close and have similar directions

When two roads are close to each other and have similar directions, it is difficult to assign traces to the right roads. In this situation, we use a fuzzy c-means clustering method to separate them. The fuzzy c-means algorithm [11] is very similar to the k-means algorithm. However, in fuzzy c-means clustering, instead of belonging completely to just one cluster, each point has a degree of belonging to each cluster.

The procedure of the fuzzy c-means clustering method involves an optimization of an objective function, that is,

$$J(U, c_1, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m d_{ij}^2 \quad (1)$$

where ci represents the ith cluster center, uij denotes the degree of belonging of jth point to ith cluster center, parameter m > 1 is a weighting exponent that determines the amount of fuzziness of the resulting classification, dij=||ci-xj|| is the Euclidean distance between jth point and ith cluster center, where xj is the jth point.

With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \quad (2)$$

The degree of belonging is related to the inverse of the distance to the cluster center, and the coefficients are normalized with parameter m so that their sum is 1.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (3)$$

The algorithm first assigns two initial random cluster centers and randomly sets initial coefficients to each point for being in the clusters. Then the algorithm computes the center for each

cluster using formula (2) and recalculates the coefficients of being in the clusters for each point using formula (3) iteratively until the algorithm has converged, that is the change of the objective function (1) between two iterations is less than a given sensitivity threshold.

The clustering result is sensitive to the initial cluster centers. In order to get better result, we set the point that is nearest to the start of the perpendicular line as one initial cluster center and set the point that is nearest to the end of the perpendicular line as the other cluster center. In this way, we get the maximum separation of the hypothesized two clusters.
First, we find roads that have neighborhood roads with similar directions by checking their neighborhood. If one road has two or more neighborhood roads with similar directions within 30 meters from it, we narrow the width of profile line according to the distance between it and its nearest neighborhood road. In this way we ensure that there are two clusters to be found. We then classify them into the following classes:

1. Road has its neighborhood road on its left.
2. Road has its neighborhood road on its right.

As described in section 4.1, after the matching method we get a series of points that traces intersect with road's perpendicular line. If the road is of one of the 2 types, we separate them into 2 clusters using fuzzy c-means algorithm. We get two cluster centers and a matrix about the degree of belonging to each cluster for each point. If the road is of type 1, traces belong to the cluster near the end of the perpendicular line are sampling traces for the road. Otherwise, traces belong to the cluster near the start of the perpendicular line are sampling traces for the road. Then we look into the degree of membership matrix and select traces whose degree of belonging for that cluster is larger than 0.5. In order to get a more reliable result, we may select traces with higher degree of membership.

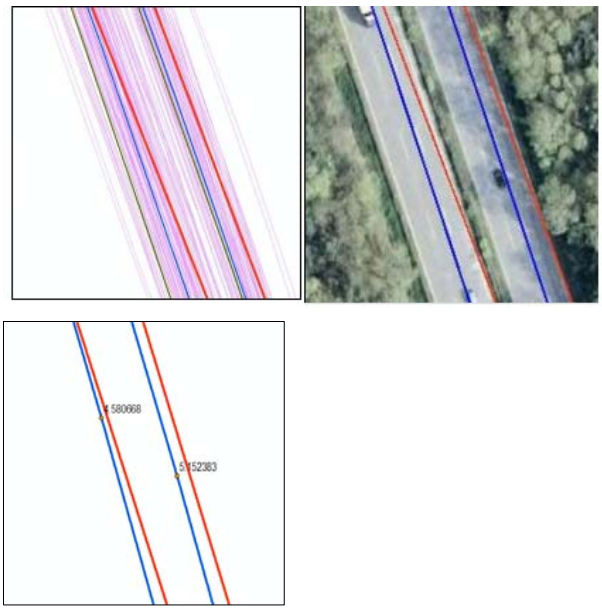
Besides the location of the intersection points, the clustering also takes the velocity of the tracks into account. The average speed of the vehicles on different roads often varies. The effect is obvious especially when separating highway exits from the highways, as the vehicles on highways exits are much slower than on the highways.

4.3 Estimating the new centerline

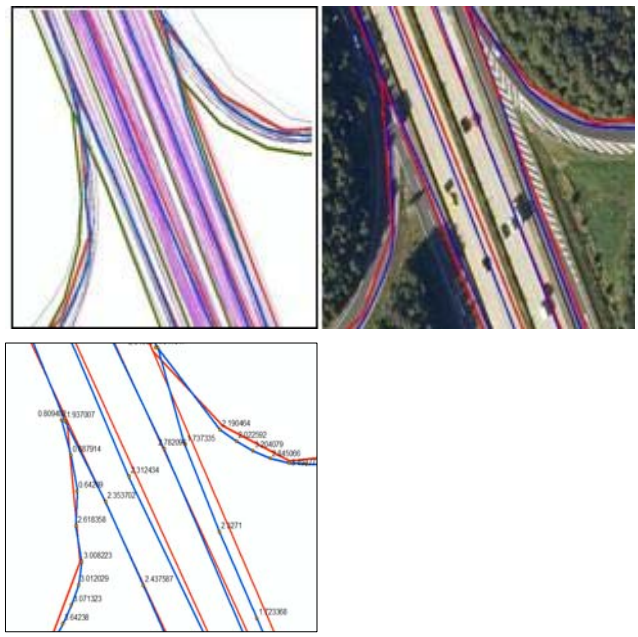
After the matching step and/or clustering step, traces are assigned to right roads. We get the intersection points of these traces with the road's profiles. Then we use a robust estimation method to select the points within 95% confidence interval, estimate the new road center vertices, and connect them to the new center line. We also add estimated standard deviations for center points to represent confidence in the points.

5. EXPERIMENTAL RESULTS

We tested our method on data from a highway area. As shown in Figure 5, the resulting roads are more accurate than the prior road map, and roads that are close and have similar direction can be separated. From Figure 7 (1) we can see that the resulting roads (in blue) are closer to TeleAtlas data, and are consistent with the centerline of the road in image data. The distance that the resulting roads move from the prior roads can reach 6 meters in some areas. As shown in Figure 7 (2), traces are separated and assigned to the right roads even when prior roads are very close to each other. The result roads are more detailed where prior roads has a high curvature. The figures also indicate the standard deviation of the centerpoints: in cases where there is only one lane, they are obviously lower than in the multi-lane case. Similarly, they are lower when a large number of GPS-traces has been used. Thus, the standard deviation both represents the accuracy of the measurement of the centerline and is an indication for the width of the road.



(1) Result where roads have different directions



(2) Result where roads are close and have similar directions

Figure 5: Experiment result: prior map is presented in red line, the result centerline is presented in blue line, and green line is the TeleAtlas road map.

In order to evaluate the result quantitatively we compared the result with a standard road map. The standard road map is from TeleAtlas dataset (in GDF-Format); it has an accuracy of 2 to

10 meters. In order to check the positional quality, we used a buffer approach as proposed by Goodchild & Hunter (1994), i.e. we evaluate the distance of the a priori road (OSM) and our result from the TeleAtlas data, which is considered of higher positional accuracy. To this end we split the result roads and the prior roads into line segments, and compare the number of line segments that are completely within 2, 5, 7 meter buffers of the TeleAtlas road map, respectively. The result is shown in Table 1. In general, the results of our methods fit better to the TeleAtlas data set than the roads from the OpenStreetMap. Considering the accuracy of GDF data in a range of 3-5 m, we can conclude that 61,7% of the centerlines derived from GPS-tracks are conform with this accuracy.

Buffer size (meter)	2	5	7
Result roads	27.4%	61.7%	73.9%
Priori roads (OSM)	14.8%	46.8%	65.8%

Table 1. Evaluation: Rates of result roads and priori roads that are within 2, 5, 7 meters buffers of the standard road map, respectively.

The result seems good when compared with TeleAtlas road map. There are, however, some cases, where the roads were wrongly reconstructed, and also roads whose position is worse than in the original road map. We analyzed these cases and found they are mainly caused by two reasons:

1. When a road does not have enough sampling GPS traces the reconstruction may be affected by its neighboring roads, which might be too far away.
2. Errors in the original road map (i.e. the prior information) may lead to errors in the result map.

6. INCREMENTAL DATA ACQUISITION

In general, there are two ways to interpret the data: on the one hand, each track can be seen as a measurement for a (static) road, where more measurements lead to an increase of accuracy of the road description. This has been described in the previous chapters, where the analysis relies on an adequate number of sample tracks.
On the other hand, the measurements can also reflect changes that have happened in the underlying phenomenon. In order to detect them, a temporal analysis of the data is necessary. In this chapter, we will outline how to make use of the incremental flow of information.
Changes in the GPS-tracks can be interpreted with respect to geometry and semantics:

- Changes in geometry: they occur in case the road has been displaced, demolished or a new road has been built. They may also occur, when there is a temporary blockage of the road (e.g. by accident or construction site), leading to the fact that cars have to change the lane to overtake the obstacle.
- Changes in attributes: road attributes refer to direction, number of lanes, traffic rules, and usage.

The changes can be identified based on an analysis of the incrementally acquired data. A new incoming data set will be analyzed with respect to its conformance with the already given data. If the differences are outside the current quality range of the road, then the information can either be considered as an outlier (error) or as new information. Thus an important issue is the differentiation and distinction of outliers in the data vs. the interpretation as a new phenomenon. This distinction will depend on the variability of the data and on the possible alternative interpretations of the new data. In this way, an optional new interpretation of the data has to be available with a certain accuracy in order to be reliably distinguished. This can be implemented using filtering techniques like Kalman or particle filtering [Thrun et al., 2005].

To this end, possible alternative interpretations of given interpretation will have to be set up and tested. This corresponds to the identification of certain temporal events. In the following, examples for alternative interpretations are given (see Figure 6):

a) If no tracks are available for certain roads any more, this is an indication that this road has ceased to exist.
b) If tracks are available at a new location, where there is no existing road yet, this is an indication for the creation of a new road.
c) if the dominant road geometry changes from a straight line to a line with a (limited) extrusion, this gives rise to a short blockage of the road which forces the drivers to overtake and use the other lane.
d) If the dominant road geometry does not continue but turns around, this is an indication for a total blockage of the road.

Figure 6: different situations where geometry of underlying feature changes (at least temporarily).

In order to detect temporally varying patterns, methods for pattern interpretation have to be implemented, which mainly are based on hypotheses about the phenomenon. Figure 7 lists some cases, e.g.,

a) in order to detect temporal variations of the usage of roads, the occupancy of cars on the roads (i.e. the GPS-points on the roads) are spatially analyzed e.g. in discrete cells or sections on the road. A temporal analysis on the number of vehicles during a day / a week / a year allows to identify these patterns.
b) The dominant direction of the traffic on the lanes can change.
c) Changed traffic rules can be detected by observing the dominant turnings at junctions. If these turnings change (with a certain statistical certainty), a change in the traffic rules can be inferred.
d) The last example shows a case where the velocity of the tracks in the vicinity of junctions has to be investigated in order to find out that a stop sign has been placed.

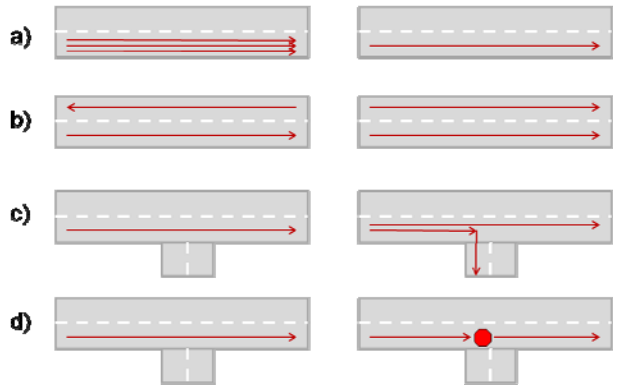


Figure 7: examples for changes in traffic rules and usage of the road.

In order to detect these events, a strategy for the interpretation has to be set up. The derivation of the road geometry in general, as well as the geometry related events (Figure 6) should in principle be applied whenever a new track is available.

On the other hand, there are events, which occur only at certain places, e.g. junctions, therefore, only those locations have to be investigated. This holds, e.g. for the detection of changes in turning restrictions (Figure 7c).

Temporal patterns related to the usage of the roads can be used for different purposes: if it is applied for the actual traffic control, the current situation is relevant and has to be collected. Furthermore, the temporal variations over time can also be included (as being done in TomTom's IQ Routes), this implies an incremental accumulation and update of the derived patterns.

For the future detailed investigations are needed in order to determine the geometric and temporal resolution necessary to reliably detect the above described events.

7. CONCLUSION AND FUTURE WORK

In this paper we have demonstrated a method for the improvement of existing road data with incoming, massive amounts of data possibly of low quality. Exploiting the potentially high amount of information compensates for the lower quality. Such an approach is of great importance both for mapping agencies who have to keep their road data sets up to date at a high rate (e.g. in Germany 3 months), but also for traffic navigation data providers who face the problem of identifying and measuring changes in their data sets. Using voluntary data, or data recorded from vehicle navigation system offers great potential if exploited in the way sketched in this paper.

We match new GPS traces with existing road information according to their distance to the road, direction and the angle between the trace and road. We use fuzzy c-means clustering method to separate traces when two roads are close and have similar direction. We also extract additional attribute information from GPS traces, such as the number of lanes.

We plan to test the approach of extracting road centerline using data from urban area, where the situation is more complicated, and make some improvement to the approach if it is needed. Furthermore, we want to extend the approach to better compensate for inaccurate prior information, e.g. by incrementally approaching the density of the GPS tracks. To this end, Kohonen Feature Nets seem to be a promising method, as employed for similar problems e.g. in Sester (2009). Due to the limited number of available tracks only general information about the center lines and the number of lanes could be derived. The future work should also include finding exact location of lanes based on the extracted information of the number of lanes.

Furthermore, we are going to implement the concepts of identifying and representing changes in the data sets through a temporal analysis of the data.

REFERENCES

Cao, L. and J. Krumm, 2009: From GPS Traces to a Routable Road Map, *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM* SIGSPATIAL *GIS 2009)*, November 4-6, 2009, Seattle, WA, pp. 3-12.

Chen, Y. and J. Krumm, 2010: Probabilistic Modeling of Traffic Lanes from GPS Traces, 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2010), November 2-5, 2010, San Jose, CA.

Davies, J.J., Beresford, A.R. & Hopper, A., 2006: Scalable, distributed, real-time map generation, *IEEE Pervasive Computing*, Vol. 5, No. 4, pp. 47-54, 2006.

Goodchild, M. F. and Hunter, G. J., 1997: A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3), p. 299-306.

Goodchild, M.F., 2007: Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2007, Vol. 2, 24-32.

Guo, D., 2008: Mining Traffic Condition from Trajectories, *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 4, pp. 256-260, 2008.

Haklay, M. and P. Weber, OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Computing, 7(4), 2008.

Sayda, F., 2005: Involving LBS users in data acquisition and update, *Proceedings of the AGILE*, Portugal, 2005.

Schroedl, S., K. Wagstaff, S. Rogers, P. Langley and C. Wilson, 2004: Mining GPS Traces for Map Refinement. *Data Mining and Knowledge Discovery*, 2004 9(1): p. 59-87.

Sester, M., 2009: Cooperative Boundary Detection in a Geosensor Network using a SOM, *ICC Chile*, CD-Rom, 2009.

Thrun, S., W. Burgard & D. Fox: Probabilistic Robotics, MIT Press, 2005.

Wikipedia. fuzzy-c-means clustering. [cited 2010]; Available from:http://en.wikipedia.org/wiki/Cluster_analysis#Fuzzy_c-means_clustering (accessed 5 May. 2010).