Integration of GPS Traces with Road Map

Lijuan Zhang
Institute of Cartography and Geoinformatics
Leibniz University of Hannover
Hannover, Germany
+49 511.762-19437
Lijuan.Zhang@ikg.uni-hannover.de

Frank Thiemann
Institute of Cartography and Geoinformatics
Leibniz University of Hannover
Hannover, Germany
+49 511.762-3724
Frank.Thiemann@ikg.uni-hannover.de

Monika Sester
Institute of Cartography and Geoinformatics
Leibniz University of Hannover
Hannover, Germany
+49 511.762-3588
Monika.Sester@ikg.uni-hannover.de

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 out of copyright road map towards a more accurate, up-to-data and detailed road map. First we profile the existing road by a sequence of perpendicular lines 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 and turning restrictions of the roads.

Categories and Subject Descriptors
I.7.5 Graphics recognition and interpretation, H.2.8. Data Mining

General Terms
Algorithms, Measurement, Experimentation.

Keywords
Integration, Incremental refinement, GPS data, Road map

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. All the above make it possible to generate road map from GPS data, and compared with 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 [7]. 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 and turn restrictions of the road. In this way, incrementally the existing information can be improved with each incoming new data set.

2. RELATED WORK
In recent years new data sources are being available like massive amounts of data collected by volunteers [5] 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 [8] 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 [9] propose to find clusters in dedicated distances, corresponding to the typical width of lanes. Cao & Krumm [1] use an approach

ACM SIGSPATIAL'10, November 2-5, 2010, San Jose, CA, USA.

based on a force model which optimizes the displacement of individual tracks towards a modeled center line. Chen & Krumm [2] 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., 2006 [3], use a raster-based – similar to occupancy grids used in robotics – in order to determine the geometry of roads. In their approach, they also include a temporal component by including a kind of fading of roads which are not regularly frequented. In this way, also abandoned roads can be detected. Thus they are able to also describe the temporal change of objects.

Guo [6] presents an approach to derive also attribute information from the GPS-tracks.

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 some attributes about the road, e.g. turning restrictions and one-way roads.

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 neighborhood 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, 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.

For our investigations we used two data sets: one containing mostly highways, and a second one from an inner city area.

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 kilometer 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 kilometer per hour. In the urban area we use 100 kilometer per hour as a threshold to calculate the speed of the traces. Figure 1shows 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: dark color indicates low 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.

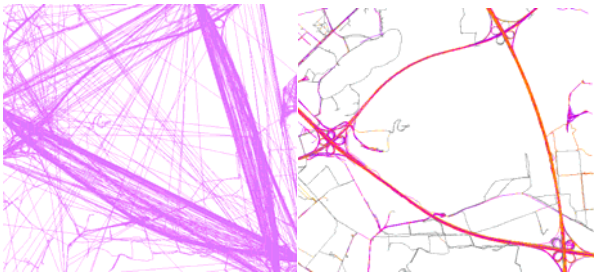
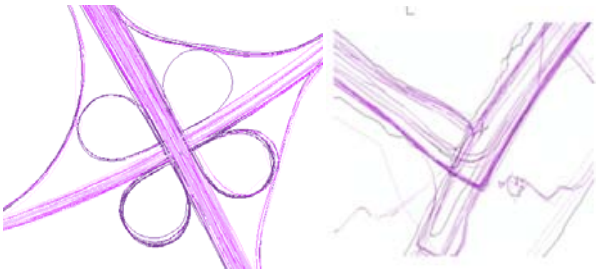


Figure 1: GPS traces in highway area, before and after preprocessing.



(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 center line, 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 center line. The whole process for the extraction of the road centerlines 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.

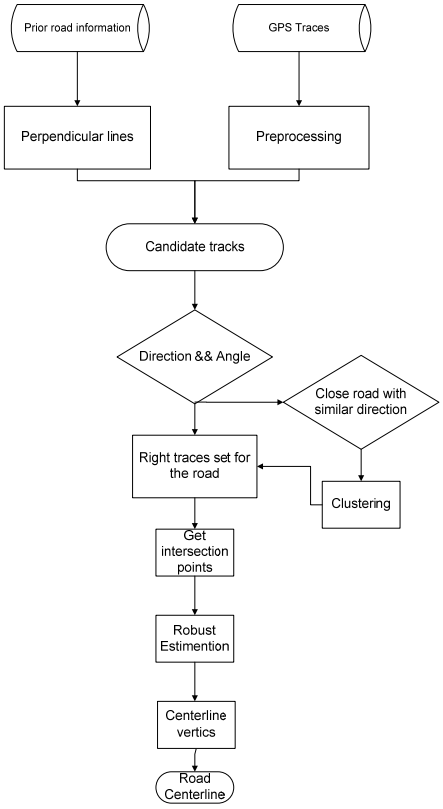


Figure 3: Work flow for extracting of road centerline.

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.

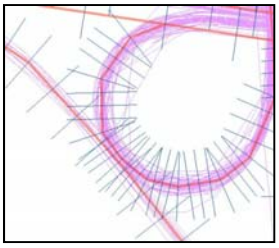


Figure 4: Getting candidate traces for the road using lines perpendicular (in green) to prior road's centerline (in red).

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.

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

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

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

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

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 1 and recalculates the coefficients of being in the clusters for each point using formula 2 iteratively until the algorithm has converged, that is the coefficients' change 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 and classify them into the following classes:

1. Road has a direction and its neighborhood road is on its left
2. Road has a direction and its neighborhood road is on its right
3. Road has no direction and its neighborhood road is on its left
4. Road has no direction and its neighborhood road is 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 4 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 1 or 3 type, 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.

4.4 Estimation of number of lanes

We also estimate the number of lanes for each road. GPS traces are expected to cluster near the center of lanes. However, as shown in Figure 6, due to the errors of the GPS data – and possibly also due to lacking number of traces – traces are not separated for each lane. This can be seen in figure 6, which shows a section of a three-lane road and the corresponding number of traces in terms of a histogram. The spread of the traces for a road can be modeled as a Gaussian distribution. Thus more sophisticated approaches like proposed by Chen & Krumm cannot be applied. We therefore consider the mean of the Gaussian distribution as the center of the road. The standard deviation σ of the Gaussian distribution can be put into a relation to the width of the road (see Figure 5Figure 6). After an analysis of our data we concluded that the width roughly corresponds to 2 σ.

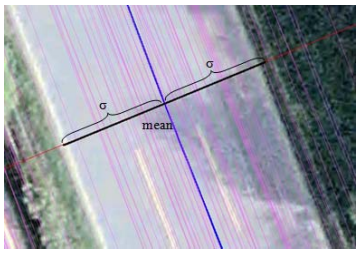


Figure 5: Standard deviation σ as a measure for the width of the road

The placement of the vehicle can be in a wider range in a multi-lane road than in a one-lane road. Therefore, the width of the road affects the spread of the traces and this can be reflected by standard deviation σ of Gaussian distribution. Analyzing our test data set we found out that if a road has more than 2 lanes, the width of each lane is about 3.5 meter, and the width of a normal one-lane road is about 5 meters. Thus we calculate the number of lanes using the following method: if the value of 2σ is smaller than 5.5, the road is a one-lane road. If 2σ is larger than 5.5, then the number of lanes is 2σ/3.5.

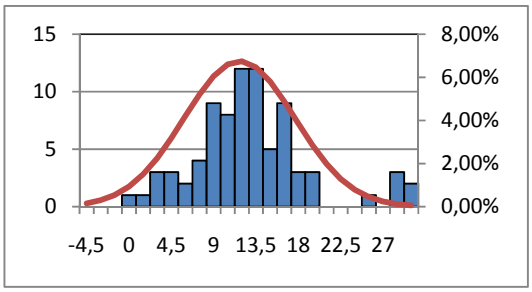


Figure 6: The distribution of GPS traces for a road can be modeled as a Gaussian distribution.

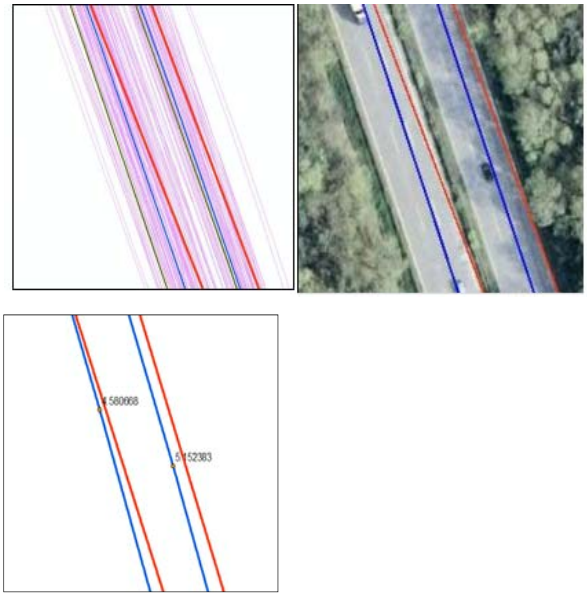
5. EXPERIMENTAL RESULTS

We tested our method on data from a highway area. As shown in Figure **7,** 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 6 (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 6 (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.

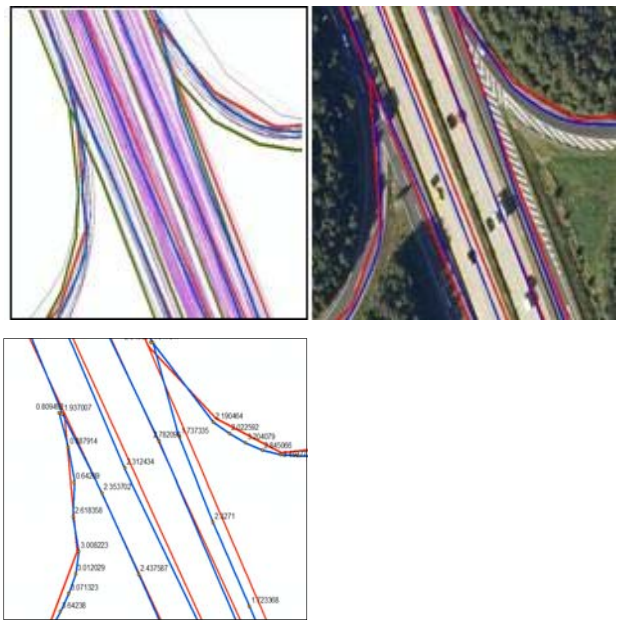
In order to evaluate the result quantitatively we compared the result with a standard road map. The standard road map is from TeleAtlas dataset (in GDF-Format); it has an accuracy of 2 to 10 meters. In order to check the positional quality, we used a buffer approach as proposed by Goodchild & Hunter, 1997 [4], 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.

The result seems good when compared with TeleAtlas road map. There are, however, wrong reconstruction, or roads that are worse than 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.



(1) Result where roads have different directions



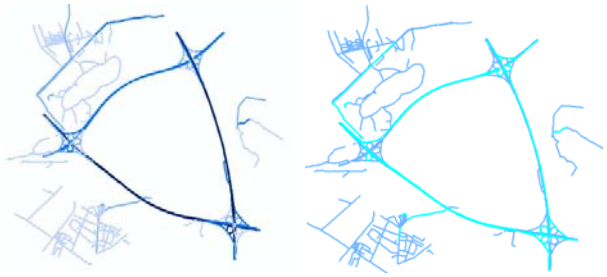
(2) Result where roads are close and have similar directions

Figure 7: 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.

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

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

We also checked the result of our method for estimating the number of lanes for 42 roads. For 25 roads the correct number was determined. As shown in Figure 9, the differences of estimated numbers and true numbers decreases noticeably when the numbers of traces used increase. Especially, when the number of traces is larger than 16, 13 roads out of 16 roads are given the correct numbers of lanes. The reason is that, if there are not enough traces, the spread of the traces cannot be modeled as Gaussian distribution. The experiment result shows that, if the number of sampling traces is larger than 16, the result of extracting number of lanes is reliable (see Figure 8).



(1) (2)

Figure 8: (1) Roads in darker blue have more sampling traces than roads in lighter blue. (2) Roads that are given correct number of lanes (highlighted lines).

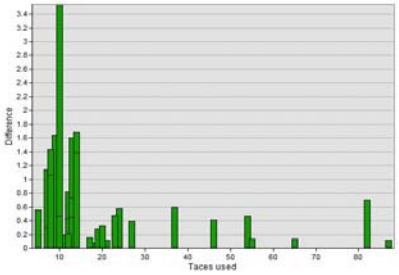


Figure 9: Difference of calculated number of lanes from the true number.

6. EXTRACTION OF ATTRIBUTE INFORMATION

Using the data it is also possible to look at additional attribute information that can be deduced from the data. E.g. the fact that some roads are only used in one direction gives rise to a one-way road. Similarly, also turning restrictions can be derived from the data sets. As the prior road map has the information of one-way or not, we use it as prior information and focus on the turn restrictions. In the following, we qualitatively analyze some junctions in order to show the potential of the data to reveal this attribute information.

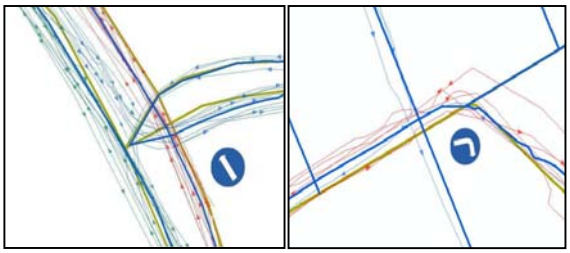
At a junction we first define which roads are possible to turn to from current road according to the connectivity and the directions of roads. Then we classify them to straight-on road, left-turn road and right-turn road for the current road.

We explore the GPS traces to analyze the road's turn restrictions. We specify the turn restrictions as the following 6 types:

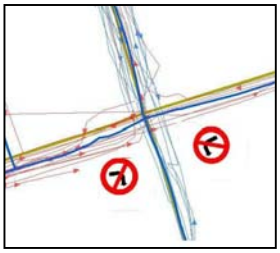
1. Only_straight_on: if all traces of the road turn to its straight-on road after the intersection.

2. Only_left_turn: if all traces of the road turn to its left-turn road after the intersection.
3. Only_right_trun: if all traces of the road turn to its right-turn road after the intersection.
4. No_straight_on: if only 2 and 3 do happen and no trace of the road continues straight-on road after the intersection.
5. No_left_turn: if 1, 2 and 3 do not happen and no trace of the road turns to its left-turn road after the intersection.
6. No_right_turn: if 1, 2 and 3 do not happen and no trace of the road turns to its right-turn road after the intersection.

In the above cases U-turns not considered. As shown in Figure 10, there are three examples for road turn restrictions. The lack of data may lead to wrong decisions about turn restriction. As the GPS data accumulating we can expect better results.



(1) Only_straight_on (2) Only_right_turn



(3) no_right_turn and no_left_turn

Figure 10: Examples of turn restrictions. GPS traces for different roads are shown in different colors with arrows indicating their direction. Brown lines are reference map and new road centerlines are represented using blue lines.

7. CONCLUSION AND FUTURE WORK

In this paper we have demonstrated a method for incremental 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. 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 number of lanes, turn restrictions of the roads.

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 [10]. 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, also investigations with respect to structural changes of the prior information has to be included, e.g. the fact that a new lane is built, that new turning instructions are introduced: the incremental process has to take possible and plausible changes into account and allow these structural changes as soon as enough current information votes for it.

8. REFERENCES

[1] Cao, L. and J. Krumm, 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.

[2] Chen, Y. and J. Krumm, Probabilistic Modeling of Traffic Lanes from GPS Traces, unpublished.

[3] 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.

[4] 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.

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

[6] Guo, D., 2008: Mining Traffic Condition from Trajectories, Fuzzy Systems and Knowledge Discovery, 2008.

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

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

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

[10] Sester, M., 2009: Cooperative Boundary Detection in a Geosensor Network using a SOM, ICC Chile, 2009.

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