An Efficient Traffic Forecasting System Based on Spatial Data and Decision Trees

Kalli Srinivasa Prasad[1] and Seelam Ramakrishna[2]
[1]Research Scholar in Computer Science, Sri Venkateswara University, India
[2]Department of Computer Science, Sri Venkateswara University, India

Abstract: *The rapid proliferation of Global Position Service (GPS) devices and mounting number of traffic monitoring systems employed by municipalities have opened the door for advanced traffic control and personalized route planning. Most state of the art traffic management and information systems focus on data analysis, and very little has been done in the sense of prediction. In this article, we devise an efficient system for the prediction of peak traffic flow using machine learning techniques. In the proposed system, the traffic flow of a locality is predicted with the aid of the geospatial data obtained from aerial images. The proposed system comprises of two significant phases: Geospatial data extraction from aerial images, and traffic flow prediction using See5.0 decision tree. Firstly, geographic information essential for traffic flow prediction are extracted from aerial images like traffic maps, using suitable image processing techniques. Subsequently, for a user query, the trained See5.0 decision tree predicts the traffic state of the intended location with relevance to the date and time specified. The experimental results portray the effectiveness of the proposed system in predicting traffic flow.*

Keywords: *Traffic flow, traffic prediction, spatial data mining, spatial data base, see5.0, decision tree algorithm.*

Received January 9, 2012; accepted May 22, 2012, published online April 4, 2013

1. Introduction

Data mining is usually defined as searching, analyzing and sifting through large amounts of data to find relationships, patterns, or any significant statistical correlation. The technical progress in computerized data acquisition and storage has resulted in the growth of vast databases. With the continuous increase and accumulation, the huge amounts of the computerized data have far exceeded human ability to completely interpret and use. In order to understand and make full use of these data repositories, a few techniques have been tried, e.g., expert system, Database Management System (DBMS), spatial data analysis, machine learning, and Artificial Intelligence (AI) [15]. Spatial Data Mining (SDM) is the process of discovering interesting, useful, non-trivial patterns information or knowledge from large spatial datasets. Extracting interesting and useful patterns from spatial datasets must be more difficult than extracting the corresponding patterns from traditional numeric or categorical data due to the complexity of spatial data types, spatial relationships, and spatial auto-correlation [29].

SDM is a new and rapidly developing area of data mining concerned with the identification of interesting spatial patterns from data stored in spatial databases and geographic information systems. Geographic Information Systems (GIS) enable capturing, storing, analyzing, and managing data and associated attributes which are spatially referenced to the Earth. GIS are used in various areas such as environmental impact assessment, urban planning, cartography, criminology, traffic analysis, etc., [13, 22]. Here, we have undertaken the process of traffic analysis based on information available in GIS. With increasing traffic volumes on urban roads, particularly in the large cities, existing roundabouts and priority-controlled junctions are being replaced with traffic signals to address capacity, efficiency and safety issues or to provide better amenity for pedestrians and cyclists [21]. Daily traffic jams reflect the fact that the capacities of the road network are not satisfied or even exceeded [2]. It is therefore crucial to investigate new technologies and alternative methods of traffic management to reduce congestion without increasing road space [1].

SDM has been shown to significantly help improving traffic safety, and has been used in many traffic related works [4]. In traffic data, traffic density patterns on hourly, weekly, and monthly scales can be obtained from density plots. Such plots identify traffic peaks, and can be of help to traffic specialists in planning routes and safety measures, as well as to individual drivers [25]. Traffic control systems for large traffic networks have attracted much attention, recently. One challenge of traffic controlling is the prediction of the traffic. Traffic flow prediction, i.e. conditionally, forecasting the traffic conditions in the network, given prevailing traffic conditions, the predicted traffic demands, and the candidate control scenarios. What we need are efficient and effective methods that are able to estimate the traffic for any point of time in the future. Traffic predictions are very important as they enable to detect potential traffic jam spots. Based on the information provided from a traffic prediction system we could initiate certain traffic control methods to avoid the traffic jams. One of the most important applications of traffic control systems is the control of road network traffic. In particular at rush-hour when the risk of the occurrence of traffic jams is very high traffic control systems would be very important [9, 28].

Intelligent Transportation Systems (ITS) has been studied and developed widely. Especially Advanced Traveler Information System (ATIS), i.e., the main part of ITS, provides the real-time traffic information to travelers, thus helping them to find alternative routes, which would reduce the delays caused by both incidents and congestions. A wide variety of techniques for the prediction of traffic volumes based on the framework of ATIS [8] have been proposed aiming at solving congestion problems. Of late, an improved solution, Dynamic Traffic Assignment (DTA) [6] has attracted increasing attention, because of its capability to process time-varying properties of traffic flow. However, these complex formulations generally lead to extremely complicated solutions and another issue is that the requirement of time-dependent Original and Destination (OD) data in prediction process [32]. In this paper, we propose a simple but efficient system for predicting peak traffic flow using image processing and machine learning techniques. The input to the proposed system is aerial images like traffic maps that state the volume of traffic in a particular locality. The proposed system makes use of the See5.0 decision tree to perform traffic flow prediction. Decision trees are generally good at processing on real-valued datasets than on images. Hence, initially, we make use of appropriate image processing techniques to prepare a training dataset by extracting essential geospatial data available in aerial images. Subsequently, for a user query, the trained See5.0 decision tree predicts the traffic state of the intended location with relevance to the date and time specified. The experimental results portray the effectiveness of the proposed system in predicting traffic flow.

The rest of the paper is organized as follows: A brief review of the recent researches related to traffic flow prediction is given in section 2. The proposed machine learning system based on spatial data for traffic flow prediction is presented in section 3. The experimental results are presented in section 4. Finally, the conclusions are summed up in section 5.

2. Related Works

Literature presents with a plentiful of approaches for traffic flow prediction using machine learning techniques. A selected few significant contributions related to the proposed research are presented below.

Zhou *et al.* [32] have applied Genetic Network Programming (GNP) to create a traffic flow prediction model for obtaining prediction rules from the past traffic data. And they proposed the spatial adjacency model for the prediction and two kinds of models for N-step prediction. Additionally, the adaptive penalty functions were adopted for the fitness function in order to alleviate the infeasible solutions containing loops in the training process. Furthermore, the sharing function was also used to avoid the premature convergence.

Sun *et al.* [23] have proposed a predictor for traffic flow forecasting, namely spatiotemporal Bayesian network predictor. Their approach incorporated all the spatial and temporal information available in a transportation network to carry their traffic flow forecasting of the current site. The Pearson correlation coefficient was adopted to rank the input variables (traffic flows) for prediction, and the best-first strategy was employed to select a subset as the cause nodes of a Bayesian network. Finally, traffic flow forecasting was performed under the criterion of Minimum Mean Square Error (MMSE). Experimental results with the urban vehicular flow data of Beijing demonstrated the effectiveness of their spatio-temporal Bayesian network predictor.

Tan *et al.* [24] have proposed an aggregation approach for traffic flow prediction that was based on the Moving Average (MA), Exponential Smoothing (ES), Autoregressive MA (ARIMA), and Neural Network (NN) models. The aggregation approach assembled information from relevant time series. The source time series was the traffic flow volume that was collected 24 h/day over several years. The predictions that resulted from the different models were used as the basis of the NN in the aggregation stage. The output of the trained NN served as the final prediction. To assess the performance of the different models, the naive, ARIMA, nonparametric regression, NN, and Data Aggregation (DA) models were applied to the prediction of a real vehicle traffic flow, from which data was collected at a data-collection point that was located on National Highway 107, Guangzhou, Guangdong, China.

Neto *et al.* [17] have presented an application of a supervised statistical learning technique called Online Support Vector machine for Regression (OL-SVR), for the prediction of short-term freeway traffic flow under both typical and atypical conditions. The OL-SVR model was compared with three well-known prediction models including Gaussian Maximum Likelihood (GML), Holt exponential smoothing, and artificial neural net models. The resultant performance comparisons suggested that GML, which relied heavily on the recurring characteristics of day-to-day traffic, performed slightly better than other models under typical traffic conditions, as demonstrated by previous studies.

Zhao *et al.* [31] have proposed the fuzzy time series method to predict short-term traffic flow. First, they proposed an improved fuzzy time series prediction model, i.e., ratio-median lengths of intervals two-factor high-order fuzzy time series. The prediction model simultaneously considered impact of many factors on the traffic flow formulation. For achieving higher prediction accuracy, the ratio-median lengths of intervals method was adopted to adaptively partition the universe of discourse of linguistic variable. Then it was used to predict the raw traffic flow data which were collected at Zizhu Bridge in Beijing. The experiment result verified that the improved fuzzy time series prediction model could achieve high prediction accuracy.

Chen *et al.* [3] have investigated the application of SOM in the representation and prediction of multi-dimensional traffic time series. First, SOMs were applied to cluster the time series and to project each multi-dimensional vector onto a two-dimensional SOM plane while preserving the topological relationships of the original data. Then, the easy visualization of the SOMs was utilized and several exploratory methods were used to investigate the physical meaning of the clusters as well as how the traffic flow vectors evolve with time. Finally, the k-Nearest Neighbor (kNN) algorithm was applied to the clustering result to perform short-term predictions of the traffic flow vectors. Analysis of real world traffic data showed the effectiveness of those methods for traffic flow predictions, for they could capture the nonlinear information of traffic flows data and predict traffic flows on multiple links simultaneously.

Nguyen *et al.* [18] have proposed an approach to traffic prediction using Ying-Yang Fuzzy Cerebellar Model Articulation Controller (YYFCMAC). Their model was motivated from the famous Chinese ancient Ying-Yang philosophy, which viewed everything as a product of conflict-harmony process between Ying and Yang. That principle was applied to find the optimal number of clusters and fuzzy sets in the fuzzification phase of the hybrid fuzzy-neural YYFCMAC network. The analyzed experiment on a set of real traffic data flow of the east-bound Pan Island Expressway (PIE) in Singapore showed the effectiveness of the YYFCMAC in universal approximation and prediction.

Dong *et al.* [5] have proposed an ARIMA model for the traffic flow prediction. The ARIMA model was trained according to the different period traffic data. Based on the different period data training, the ARIMA model was refined more accuracy. The experiments showed that the ARIMA model trained by the time-oriented data could reach a better result than the non time-oriented data trained model. Yin *et al.* [30] have developed a Fuzzy-Neural Model (FNM) to predict the traffic flows in an urban street network, which considered a major element in the responsive urban traffic control systems. The FNM consisted of two modules: a Gate Network (GN) and an Expert Network (EN). The GN classified the input data into a number of clusters using a fuzzy approach, and the EN specified the input-output relationship as in a conventional neural network approach. An online rolling training procedure was proposed to train the FNM, which enhanced its predictive power through adaptive adjustments of the model coefficients in response to the real-time traffic conditions. Both simulation and real observation data were used to demonstrative the effectiveness of their method.

Min *et al.* [16] have presented a method which provided a complete description of the most important spatio-temporal interactions in a road network while maintaining the estimatability of the model. It improved upon existing methods proposed in the area and provided high accuracy on both urban and expressway roads. The accuracy exceeded that of other published works on 15-minute data, and could achieve very good accuracy on the more volatile 5-minute data. In addition, accuracy remained very good up to 12 time periods into the future.

3. Proposed Machine Learning System Based on Spatial Data for Traffic Flow Prediction

Geospatial data are emerging as an increasingly significant component in decision making processes and planning efforts across a wide range of industries and information sectors. Geospatial data, also termed as geographic information or as spatial data based on the context, can be defined as data that explain features on the earth Figure 1. Typically, datasets like transportation networks, property boundaries, coastlines, aerial imagery, or terrain models can all be well thought-out to be geospatial data. In this paper, we describe a machine learning system that performs traffic flow prediction based on geospatial data obtained from aerial images portraying traffic information. The dataset used in the proposed system corresponds to transportation networks that contain the state of traffic in different regions of the locality.

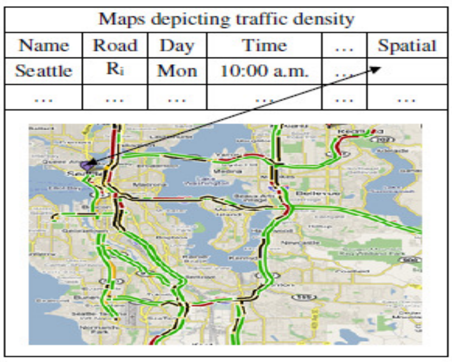


Figure 1. Sample traffic map with the associated geospatial data.

The proposed system comprises of two significant phases:

1. Geospatial data extraction from aerial images.
2. Decision module for traffic flow prediction using See5.0 decision tree.

3.1. Geospatial Data Extraction from Aerial Images

The necessity for accurate geospatial traffic data is rising speedily so, as to proffer effectual traffic flow prediction in urban areas. Geospatial databases hold a variety of man-made objects among which roads are of special significance as they are employed in an assortment of applications such as car navigation, transport and fire services. Since, their extraction from images is expensive and time-consuming, automation seem as a promising solution to this dilemma. The problem with automatic data extraction is generally, because of the complex content of aerial images. But, the timely road information can serve a world of good to the decision makers for urban planning, traffic management and more [27]. The work presented in this paper focuses mainly on the prediction of traffic flow. The steps involved in the process of extracting the road information from the images are:

- Conversion of color image into grayscale image.
- Binarization.
- Morphological operators.
- Extraction of traffic information.

The block diagram of the steps involved in geospatial data extraction from aerial images is depicted in Figure 2.

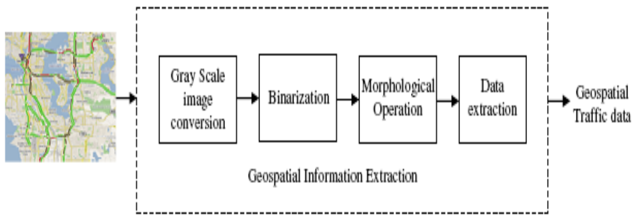


Figure 2. Block diagram of the steps involved in geospatial data extraction from aerial images.

3.1.1. Conversion of Color Image into Grayscale Image

The original images (maps depicting traffic information) comprise of an associated Red-Green-Blue (RGB) color map by "R" and "G" for "red" and "green" with R corresponding to the Cy5 and G to Cy3 respectively. In order to extract significant information from the images and address the spots, it is necessary to first convert them into grayscale images. It is only an auxiliary step since, after the spots are located; the original red and green channels can be used to extract the true intensities. In order to get the grayscale images, the RGB color model is converted to a YIQ (luminance-hue saturation) model. This model has the advantage that decouples luminance and chromaticity, codifying in different channels grayscale and color data. To obtain the grayscale information, the I and Q components are set to zero. The Y component is obtained by means of the weighted sum of the R, G and B channels, as described in the following equation [19]:

$$Y=0.299R+0.58G+0.114B \quad (1)$$

3.1.2. Binarization

The next step in geospatial data extraction is the process of binarization, which converts the grayscale image into a binary image so, as to improve the contrast of the road segments from rest of the image. The proposed system makes use the global binarization technique of Otsu [19]. Global binarization is a method that has single threshold T for the entire image. The pixels with gray level greater than T are labeled as '1', while the others are labeled as '0' [14]. The binarization process involves [26]:

1. Investigating the gray-level value of each pixel in the enhanced image.
2. If the value is greater than the global threshold, then the pixel value is set to a binary value one; otherwise, it is set to zero.

3.1.3. Morphological Operators

Then, the binary morphological operators are applied on the binarized image. The morphological operators are applied mainly for the purpose of removing any of the obstacles and noise from the image. Moreover, the unwanted spurs, bridges and line breaks are removed by these operators. Subsequently, thinning process is performed to reduce the thickness of the lines so that the lines (roads) are only represented except the other regions of the image Thinning can be defined as a morphological operation that efficiently erodes away the foreground pixels till they become one pixel wide. It aids in the removal of redundant pixels till the ridges become one pixel wide [7, 12].

3.1.4. Extraction of Traffic Information

The 'regionprops', a function in MATLAB's Image Processing Toolbox is utilized in the proposed system for extraction of the roads from the images:

$$STATS = regionprops\ (BW, properties) \quad (2)$$

The 'regionprops' function measures the set of properties for each connected component (object) in the binary image, BW. The image BW is generally, a logical array of some dimension. The aforesaid set of processes extract the geospatial traffic information that can aid in the subsequent procedures involved in predicting traffic flow. These set of historical information decide on the future traffic density of a given place, in a given date and time.

3.2. Proposed Decision Module for Traffic Flow Prediction-Decision Tree

The proposed system for traffic flow prediction employs decision trees to decide on the future traffic state of the intended place. The proposed system takes as input a user query requesting for traffic information. The user query comprises of the locality attributes (XY coordinates), the day and the time during which the travel is planned. The decision module in the proposed approach first fetches records relevant to the user query from the collected historical information. Subsequently, the fetched records are fed to the trained See5.0 decision tree for classification. The See5.0 decision tree classifies the relevant data to their appropriate class labels. Finally, the class occurring maximum number of times for the relevant records is chosen as the prediction result. The block diagram of the decision module involved in the proposed system is depicted in Figure 3.

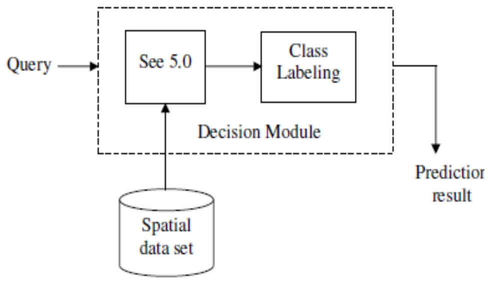


Figure 3. Block diagram of the proposed decision module for traffic flow prediction.

Regarding decision trees, they are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). The mostly commonly used are the classification trees that are usually represented graphically as hierarchical structures, making them easier to interpret than other techniques. Classification trees are used to classify an object or an instance such as insurant to a predefined set of classes such as risky/non-risky based on their attributes values such as age or gender. Classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine.

Decision tree generally consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an “internal” or “test” node. All other nodes are called “leaves” (also, known as “terminal” or “decision” nodes). In the decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector affinity vector indicating the probability of the target attribute having a certain value.

There have a variety of algorithms for building decision trees that share the desirable quality of interpretability. A well known and frequently used over the years is C4.5 (or improved, but commercial version See5/C5.0). In the proposed system for traffic flow prediction, we make use of the decision tree algorithm, See 5.0. The See 5.0 can be seen just as an extension improvement of C4.5, with basic concepts holding for either of the algorithms. So, an illustration of the C4.5 algorithm could give a better picture of the working of the See5.0 algorithm [10, 11, 33]. See 5.0 allows a range of data samples and can be loaded with user-specified data from a file of comma separated values. Test data, misclassification costs and a range of options can be specified before a decision tree is constructed. These include pruning, winnowing, boosting and the ability to set fuzzy thresholds. The decision tree is generated as an ASCII representation with misclassification rates. The tree can then be examined using new user-defined instances for its prediction accuracy and can also be cross-examined with instances from the training or test sets. Output can also be converted to a set of rules.

- *Improvements in See5.0*

See5.0 offers the improvements on C4.5 are [20]:

- Speed-See5.0 is appreciably faster than C4.5 (several orders of magnitude)
- Memory usage-See5.0 is comparatively more memory efficient than C4.5
- Smaller decision trees-See5.0 gets analogous results to C4.5 with significantly smaller decision trees.
- Support for boosting-Based on the research of Freund and Schapire, this is an exciting new development that has no counterpart in C4.5. Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy. C5.0 uses a proprietary variant of boosting that is less affected by noise, thereby partly overcoming the limitations of C4.5. C5.0 supports boosting with any number of trials. Naturally, it takes longer to produce boosted classifiers, but the results can justify the additional computation. Boosting should always be tried when peak predictive accuracy is required, especially when unboosted classifiers are already quite accurate.

Boosting works by creating multiple training sets from one training set. Each item in the training set is assigned a weight. The weight indicates the importance of this item to the classification. A classifier is constructed for each combination of weights used. Thus, multiple classifiers are actually constructed. When C5.0 performs a classification, each classifier is assigned a vote, voting is performed, and the target tuple is assigned to the class with the most number of votes.

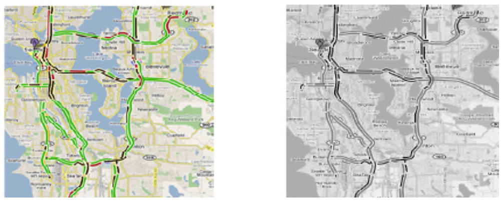
- Weighting-See5 allows you to weight different attributes and misclassification types. C5.0 incorporates several new facilities such as variable misclassification costs. In C4.5, all errors are treated as equal, but in practical applications some classification errors are more serious than others. C5.0 allows a separate cost to be defined for each predicted/actual class pair; if this option is used, C5.0 then constructs classifiers to minimize expected misclassification costs rather than error rates. C5.0 has provision for a case weight attribute that quantifies the importance of each case; if this appears, C5.0 attempts to minimize the weighted predictive error rate. C5.0 has several new data types. In addition, to those available in C4.5, including dates, times, timestamps, ordered discrete attributes, and case labels. In addition, to missing values, C5.0 allows values to be noted as not applicable. Further, C5.0 provides facilities for defining new attributes as functions of other attributes.
- Winnowing-See5 automatically winnows the data to help reduce noise. C5.0 can automatically winnow the attributes before a classifier is constructed, discarding those that appear to be only marginally relevant. For high-dimensional applications, winnowing can lead to smaller classifiers and higher predictive accuracy, and can even reduce the time required to generate rule sets.

In pseudo-code the algorithm looks like this:

- Check for base cases.
- For each attribute a.
- Find the normalized information gain from splitting on a.
- Let a_best be the attribute with the highest normalized.
- Information gain.
- Create a decision node that splits on a_best.
- Recurse on the sub-lists obtained by splitting on a_best and add those nodes as children of node.

4. Experimental Results

In this section, we have presented the experimental results of the proposed system for traffic flow prediction. The implementation setup of the proposed system comprises of, Geospatial data extraction from aerial images using MATLAB (Matlab 7.8) and; decision programmed in Java (jdk 1.6). The input to the proposed system is traffic maps that depict the traffic densities of distinct locations at different periods of time. The proposed system has been validated on traffic maps obtained from publicly available datasets. First, the traffic maps are subjected to a set of image processing techniques for geospatial data extraction. The input image and the intermediate results obtained during various steps such as Grayscale image conversion, Binarization, Morphological operations and Blob analysis are portrayed in Figures 4 and 5.



a) Input image. b) Grayscale converted image.

Figure 4. The input image and the intermediate results.

a) Binarized image. b) Morphological operations. c) Information extraction from individual roads.

Figure 5. The intermediate results obtained during various steps.

The working of the proposed decision module can well be described by means of a sample query image. Given in Figure 6 is the query image with locality attributes (102, 102) and with it the user provides auxiliary information namely day Monday and time (10:00 am).

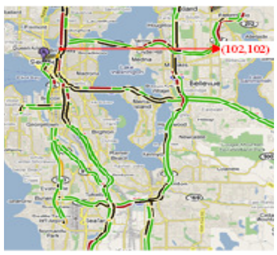


Figure 6. Query image depicting locality attributes.

Here, we briefly describe the explanatory variables in a train or test file portrayed in Table 1: Day is the day of the week from Sunday to Saturday, Time is the time when the user is expected to use the road, R_i is the road index, Coordinates X and Y are the attributes that describe the locality within the image, R is the intensity of Red, G is the intensity of green, B is the intensity of blue, Class is the label assigned to the test

Table 1. Train/ test file with prediction result.

Day	Time	Road	Coordinate X	Coordinate Y	R	G	B	Class label	Prediction Result
MON	10:00 a.m.	R_i	102	102	157	2	0	High Congestion	Moderate Congestion
MON	10:00 a.m.	R_i	102	102	253	204	1	Moderate Congestion	
MON	10:00 a.m.	R_i	102	102	2	1	0	Very High Congestion	
MON	10:00 a.m.	R_i	102	102	254	202	2	Moderate Congestion	
MON	10:00 a.m.	R_i	102	102	153	2	1	High Congestion	
MON	10:00 a.m.	R_i	102	102	255	206	0	Moderate Congestion	

case and prediction result is the final class label obtained by majority voting. Traffic Flow Graph of a Time Interval in a Day is depicted in Figures 7, 8 and 9, respectively.

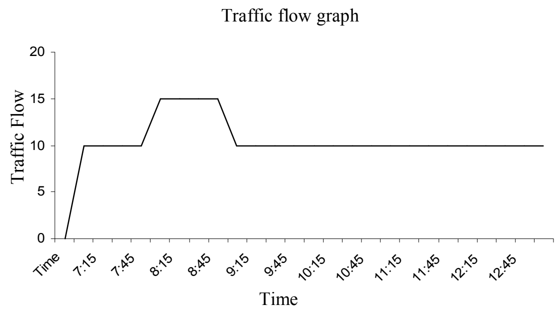


Figure 7. Traffic flow graph of a time interval (07:00 to 13:00, Friday).

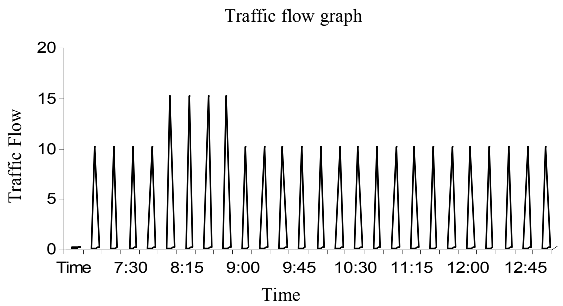


Figure 8. Traffic flow graph of a time interval (07:00 to 13:00, Friday).

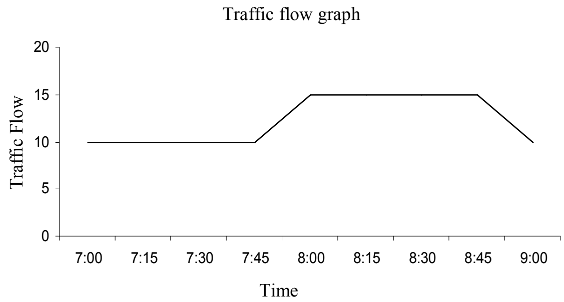


Figure 9. Traffic flow graph of a time interval (07:00 to 09:00, Friday).

5. Conclusions

In this paper, we have presented a machine learning system that performs effectual traffic flow prediction based on geospatial data obtained from aerial images. The aerial images employed in the proposed approach are traffic maps that describe the traffic state of a particular locality. The two significant phases that make-up the proposed system are, Extraction of Geospatial data from aerial images and; prediction of traffic flow using See5.0 decision tree. To start with, the proposed system has employed suitable image processing techniques for extracting geographic information essential for traffic flow prediction from aerial images such as traffic maps. Then, thse trained See5.0 decision tree has been made use of to predict the traffic state of the intended location with relevance to user query. The efficiency of the proposed system in predicting traffic flow has been demonstrated via the experimental results obtained.

References

[1] Australia M., "Review of Urban Congestion Trends, Impacts and Solutions, *Technical Report,* Competition and Regulation Working Group, 2006.

[2] Bazzan A., Wahle J., and Klügl F., "Agents in Traffic Modelling-From Reactive to Social Behaviour," *in Proceedings of the 23rd Annual German Conference on Artificial Intelligence Bonn*, Germany, vol. 1701, pp. 303-306, 1999.

[3] Chen Y., Zhang Y., and Hu J., "Multi-Dimensional Traffic Flow Time Series Analysis with Self-Organizing Maps," *Tsinghua Science and Technology*, vol. 13, no. 2, pp. 220-228, 2008.

[4] Chong M., Abraham A., and Paprzycki M., "Traffic Accident Data Analysis using Machine Learning Paradigms," *Informatica: An International Journal of Computing and Informatics*, vol. 29, no. 1, pp. 89-98, 2005.

[5] Dong H., Jia L., Sun X., Li C., and Qin Y., "Road Traffic Flow Prediction with a Time-Oriented ARIMA Model," *in Proceedings of the 5th International Joint Conference on INC*, Korea, pp. 1649-1652, 2009.

[6] Joksimovic D., Bliemler M., and Bovy P., "Optimal Toll Design Problem in Dynamic Traffic Networks with Joint Route and Departure Time Choice," *Transportation Research Board of the National Academies*, Washington, USA, vol. 1923, pp. 61-72, 2006.

[7] Kaur M., Singh M., Girdhar A., and Sandhu P., "Fingerprint Verification System using Minutiae Extraction Technique," *in Proceedings of World*

Academy of Science, Engineering and Technology, vol. 36, pp. 1-6, 2008.

[8] Kaysi I., Akiva M., and Koutsopoulos H., *Integrated Approach to Vehicle Routing and Congestion Predictions for Real-time Driver Guidance*, Transportation Research Board, pp. 66-74, 1993.

[9] Kriegel H., Renz M., Schubert M., and Zuefle A., "Statistical Density Prediction in Traffic Networks," *in Proceedings of the 8th SIAM Conference on Data Mining*, USA, pp. 692-703, 2008.

[10] Kohavi R. and Quinlan R., "Decision Tree Discovery," *in Handbook of Data Mining and Knowledge Discovery*, pp. 267-276, 1999.

[11] Kusrini and Hartati S., "Implementation of C4.5 Algorithm to Evaluate the Cancellation Possibility of New Student Applicants at Stmik Amikom Yogyakarta," *in Proceedings of the International Conference on Electrical Engineering and Informatics*, Indonesia, pp. 17-19, 2007.

[12] Lam L., Lee S., and Suen C., "Thinning Methodologies-A Comprehensive Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 869-885, 1992.

[13] Lavrac N., Jesenovec D., Trdin N., and Kosta N., "Mining Spatio-Temporal Data of Traffic Accidents and Spatial Pattern Visualization," *Metodološki Zvezki*, vol. 5, no. 1, pp. 45-63, 2008.

[14] Lee B., Park K., Kang H., Kim H., and Kim C., "Adaptive Local Binarization Method for Recognition of Vehicle License Plates," *in Proceedings of the 10th International Workshop on Combinatorial Image Analysis*, New Zealand, vol. 3322, pp. 646-655, 2004.

[15] Li D. and Wang S., "Concepts, Principles and Applications of Spatial Data Mining and Knowledge Discovery," *in Proceedings of ISSTM 2005*, China, pp. 27-29, 2005.

[16] Min W., Wynter L., and Amemiya Y., "Road Traffic Prediction with Spatio-Temporal Correlations," *IBM Research Report*, Thomas J. Watson Research Center, New York, USA, 2007.

[17] Neto M., Jeong Y., Jeong M., and Han L., "Online-SVR for Short-term Traffic Flow Prediction under Typical and Atypical Traffic Conditions," *Expert Systems with Applications: an International Journal*, vol. 36, no. 3, pp. 6164-6173, 2009.

[18] Nguyen M., Shi D., Quek C., and Ng G., "Traffic Prediction using Ying-Yang Fuzzy Cerebellar Model Articulation Controller," *in Proceedings of the 18th International Conference on Pattern Recognition*, Hong Kong, vol. 3, pp. 258-261, 2006.

[19] Otsu N., "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on System, Man, Cyber*, vol. 9, pp. 62- 66, 1979.

[20] Quinlan J., "Improved Use of Continuous Attributes in c4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

[21] Roozenburg A. and Turner S., "Accident Prediction Models for Signalised Intersections," available at: https://www.ipenz.org.nz/ipenztg/papers/2005/03_Roozenburg.pdf, last visited 2004.

[22] Shekhar S., Zhang P., Huang Y., and Vatsavai R., *Trends in Spatial Data Mining*, MIT Press, pp. 1-24, 2004.

[23] Sun S., Zhang C., and Zhang Y., "Traffic Flow Forecasting using a Spatio-Temporal Bayesian Network Predictor," *in Proceedings of the 15th International Conference*, Poland, vol. 3697, pp. 273-278, 2005.

[24] Tan M., Wong S., Xu J., Guan Z., and Zhang P., "An Aggregation Approach to Short-Term Traffic Flow Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 60-69, 2009.

[25] Tesema T., Abraham A., and Grosan C., "Rule Mining and Classification of Road Traffic Accidents using Adaptive Regression Trees," *the International Journal of Simulation: Systems, Science & Technology*, vol. 6, no. 10, pp. 80-94, 2005.

[26] Thai R., "Fingerprint Image Enhancement and Minutiae Extraction," available at: http://www.peterkovesi.com/studentprojects/raymondthai/RaymondThai.pdf, last visited 2003.

[27] Wang Y., Tian Y., Tai X., and Shu L., "Extraction of Main Urban Roads from High Resolution Satellite Images by Machine Learning," *in Proceedings of the 7th Asian Conference on Computer Vision*, India, vol. 3851, pp. 236-245, 2006.

[28] Wikipedia., "C4.5 Algorithm," available at: http://en.wikipedia.org/wiki/C4.5_algorithm, last visited 2013.

[29] Wu X., Chuin H., and Lau., "Traffic Prediction in Intelligent Vehicle Highway Systems," *in Proceedings of International Conference System Science and System Engineering*, Beijing, China, pp. 424-429, 1998.

[30] Yang T., Bai P., and Gong Y., "Spatial Data Mining Features between General Data Mining, Ettandgrs," *in Proceedings of International Workshop on Education Technology and Training & International Workshop on Geoscience and Remote Sensing*, Shanghai, vol. 2, pp. 541-544, 2008.

[31] Yin H., Wong S., Xu J., and Wong C., "Urban Traffic Flow Prediction using a Fuzzy-Neural Approach," *Transportation Research Part C:*

Emerging Technologies, vol. 10, no. 2, pp. 85-98, 2002.

[32] Zhao L. and Wang F., "Short-Term Traffic Flow Prediction based on Ratio-Median Lengths of Intervals Two-Factors High-Order Fuzzy Time Series," *in Proceedings of IEEE International Conference on Vehicular Electronics and Safety*, Beijing, China, pp. 1-7, 2007.

[33] Zhou H., Mabu S., Wei W., Shimada K., and Hirasawa K., "Traffic Flow Prediction with Genetic Network Programming," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 6, pp. 713-725, 2009.

Kalli Srinivasa Prasad has completed his MSc (Tech), MSc, MS (Software Systems), PGDCS. He is currently pursuing his PhD degree in the field of data mining at Sri Venkateswara University, India. He has published ten research papers in international journals. He has also attend three national conferences.

Seelam Ramakrishna is currently working as a professor in the Department of Computer Science, College of Commerce, Management & Computer Sciences in Sri Venkateswara University, India. He has completed his MSc, MPhil, PhD, M.Tech (IT). He is specialized in fluid dynamics and theoretical computer science. His area of research includes artificial intelligence, data mining and computer networks. He has an experience of 26 years in teaching field. He has published 44 research papers in national & international journals. He has also attended 13 national conferences and 3 international conferences. He has guided 17 PhD scholars and 18 MPhil Scholars.